International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 511-523 March 2026-Sept 2026 www.csjournals.com

Paddy yield prediction under modernization approaches using ML based on agronomic and stage wise climatic features: Comparative study on regression models.

Reetu, Dr. Sahil Saharan, Dr Anupam Bhatiya
MCA Student, Asst Professor, Associate Professor
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

The accuracy of estimation in paddy (rice) production is crucial in agricultural planning, allocation of resources and management of food security, and current estimation methods often rely on experience and have difficulty in quantifying the interactions between growers' decisions and weather data, which are mostly linear. A supervised machine learning framework is developed to predict the yield of paddy based on data collected from 2,789 paddy cultivations spread over six agricultural blocks with each block having 45 attributes covering land preparation, seeding rate, fertilizer application (DAP, Urea, Potash, micronutrient), pesticide or weedicide use as well as observation of weather parameters (rainfall, ante irrigation, temperate, wind speed, relative humidity) of a particular cultivation block, taken at four stages of crop growth. After cleaning this dataset (2,338 records) was used for training and the six regression algorithms are evaluated with Mean Absolute Error, S the same data set. A Random Forest model with the optimal parameters determined by GridSearchCV hyperparameter optimization and a five-fold cross validation yielded the best performance with R^2 coefficient of 0.99 and RMSE very close to 910 kg, followed by feature-importance analysis which reveals the seed rate, DAP dosage, micronutrient timing, and the nursery area as the major yield determinants. The results also highlight the fact that the ensemble tree-based regressors significantly outperform linear and kernel based methods and distance methods on heterogeneous agronomic climatic data, and the obtained model can be an effective decision support tool for farmers and agricultural planners by maximizing the use of inputs and improve the productivity of the rice crop.

Keywords—paddy yield prediction; machine learning regression; random forest; precision agriculture; feature importance; ensemble learning; hyperparameter tuning

I. INTRODUCTION

Nearly half of the population of India is employed in the agricultural sector, and it contributes a significant portion to India's GDP. Rice, mostly grown as paddy, is the most important staple for a majority of the people, and large gains in yield forecasting or input management due to slightly better input design can make a difference to millions of the smaller farming population. The cultivation process is, however, defined by a complex matrix of agronomic decisions made when preparing the land, establishing the crop, applying fertilizer, pesticides and weedicides and during the growth stages of the crop and climatic exposures - rain, temperature, wind and humidity - throughout these time periods.

Traditional farm and field officers strategies have relied mainly on "field" experience and general advices to plan inputs in this context. These heuristics do not usually explain the non-linear, local relationships between the different variables that actually determine the biomass harvest, and provide only little quantitative advice on what variables are important to focus on depending on the conditions. The increasing digitalization of agriculture's data alone, combined with recent developments in statistical learning, allows to do away with the subjectivity inherent in the calculation of expected yield and to instead build models that learn from past data the patterns that determine yields.

This paper examines the suitability of supervised regression techniques for modelling the yield of paddy from a combination of the management information provided by the growers under these factors and the observations of the various stages of weather, and examines which of these variables are the most important in predicting paddy yield. All six regression paradigms are learned and evaluated under the same pre-processing and evaluation procedures and the best model is then analyzed by feature-importance and residual analysis to transform the model's prediction into useful agronomic advice.

II. LITERATURE REVIEW

Since the last decade, there has been gradual growth in the interest to implement machine learning for crop-yield estimation as agricultural and meteorological records are being digitized for better access.With the increasing digitization of the records across the fields of agriculture and meteorology, there has been an increasing interest to implement machine learning for crop-yield estimation in the past decade. An overall finding from all this research is that ensemble tree-based learners, especially the Random Forest techniques, are generally better classifiers than simple, linear estimators, when the goal is to determine the outcomes of many interacting agronomic and environmental variables. Breiman's original development of Random Forest introduced the idea of feature randomisation and bootstrap aggregation to further research properties: overfitting resistance, which were increasingly adapted by later research to bootstrap specific tasks like yield modelling with mixed numerical and categorical predictors [1].

Another stream of studies has focused on how climatic variables affect rice productivity, indicating that the effects of precipitation amount and pattern, and also of climatic extremes experienced at critical crop reproductive stages have a disproportionate impact on productivity compared to the corresponding seasonal means. The stage-wise capture of the weather data (in vegetative, tillering, reproductive and ripening windows) compared to their aggregation based on the whole season has been a motivating factor because of the fact that the aggregation sometimes smoothens weather data and hides the seasonality of that weather event.

Studies on agronomic management have also focused on seed rate and when and how much of nitrogen, phosphorous and potassium fertilizers are applied, and have used correlation or importance-ranking techniques to compare the influence of the various attributes with climatic ones, often. Due to its sequential residual correction mechanism, boosting-based ensemble (e.g., Gradient Boosting (GB) and XGBoost) was shown to match or outperform Random Forest accuracy in a number of agriculture related studies [2] and [5]. SVR has also been studied, but it is universally reported to be sensitive to the scale of features, and the

performance may suffer if there are numerous dimensions with different scales in the data set, as noted in [4] and was observed during the present study.

K-Nearest Neighbours has also been used to produce estimation, which has been reported to be very strongly dependent on the number of neighbours required (K) and importantly, on the prior standardization of features. In this literature, cross-validation and an exhaustive tuning of hyperparameters performed as a safeguard against overfitting is generally recommended, notably for a large number of tunable hyperparameters in case of ensemble models. Such studies have been made much easier to reproduce and extend with widely used open-source tooling like the scikit-learn library and the pandas and NumPy libraries for data manipulation [3], [6]. The table summarizes the main themes which were highlighted in the review and their relevance in the present work.

Theme	Reported Finding	Relevance Here
Algorithm choice	Ensemble trees outperform linear/kernel models on agri-data	Random Forest selected as final model
Climatic timing	Stage-wise rainfall/temperature matter more than seasonal means	Four growth-stage weather blocks used
Fertilizer dosage	N-P-K timing strongly affects productivity	DAP, Urea, Potash, micronutrients included
Validation practice	CV + tuning improves generalization	5-fold CV and GridSearchCV applied
SVR behaviour	Sensitive to scaling, often underperforms	Confirmed by negative R^2 obtained

TABLE I. SUMMARY OF REVIEWED LITERATURE AND RELEVANCE TO THIS STUDY

III. RESEARCH GAP

While previous research demonstrates the overall usefulness of ensemble regressors for agricultural yield estimation, it fails to consider some factors identified in the reviewed literature, and directly targeted in this study.

First, much of the published literature is based on aggregated data on a district or block level instead of data collected from the field that can reveal individual cultivations and weather exposure differences according to growth stages. Secondly, studies rarely compare multiple regression paradigms over a long list of pre-processing / validation methods.Second, comparative studies that evaluate 6+ regression paradigm methods under a single, consistent pre-processing / validation pipeline are rather rare; instead, 2-3 algorithm methods are typically compared. Third, few models combine detailed information on agronomic inputs

(seed rate, fertiliser staged doses and timing, pesticide or weedicide inputs) with stage-wise climatic series within a single predictive model; instead, input timing (both stages and agronomic inputs) between the two variables are separated. Fourth, rigorous robustness checking (that is, checking by means of k-fold cross validation and exhaustive hyperparameter search) is seldom undertaken and questions remain regarding the generalizability of published accuracy figures. Lastly, little effort has been devoted to specifically prioritising agronomic inputs on climate, information directly relevant for extension officers' input prioritisation advice to farmers in the context of paddy cultivation in India.

In order to achieve this, this study aims to create a unified field-level, stage-wise dataset by preaching agronomic and climatic predictors, compare the performance of six regression algorithms using the same conditions and validate the chosen model using two approaches—cross validation and hyperparameter optimization—and provide an explicit ranked list of the most explanatory factors to predict paddy yield.

IV. OBJECTIVES

1.To compile and pre process field level agricultural data for paddy cultivation along with the climatic data of different stages.

2.To perform exploratory data analysis to understand the distributions of the features, look for anomalies and review the relationships between feature variables.

3.To train and compare 6 regression algorithms (Linear Regression, Decision Tree, Random Forest, Support Vector Regression, K-Nearest Neighbours, and Gradient Boosting) for prediction of yield of paddy crop.

4.Calculate the model performance measures: MAE, MSE, RMSE and R2, under the same train/test regime.

5.To perform model-based feature importance and rank the different agronomic and climatic factors most predictive of yield.

6.To use the best model that has been optimized using GridsearchCV, and to check the model reliability using 5-fold cross validation.

V. PROPOSED METHODOLOGY

A. Dataset

The study is based on the 2,789 cultivation records obtained from six agricultural blocks with each attribute representing a characterization of the cultivation, and each block having 45 attributes. The data cover land & cultivation information (hectares, variety, soil type, nursery type & area), agronomic inputs (seed rate, land-preparation quantity, DAP quantity, weedicide quantity, minimum and maximum temperature, wind speed and wind direction, relative humidity, before and after sunrise), and pest & residue management (pesticide quantity, relative quantity of trash), where these data series are recorded for four crop-growth periods (Days 1 – 30, 31 – 60, 61 – 90, and 91 – 120): rainfall, antecedent irrigation, minimum and maximum temperature, wind speed & direction, and relative humidity, before

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 511-523 March 2026-Sept 2026 www.csjournals.com

and after sunrise. The regression target is the yield of harvested paddy which varies from 5,410 kg - 38,814 kg with an average of 22,610 kg.

B. Pre-processing

The process of data preparation was conducted in 4 steps. There were 2,338 unique records after duplicate screening, otherwise the models would have been biased toward over-represented observations. No nulls were detected, however the imputation process used a "median for numeric" and a "mode for categorical" approach to ensure that the data could work with the imputation process in the event that nulls appeared in sources field in the future. All the input features were scaled to zero mean and unit variance, with special importance on the items that require to be scaled as (soil type, nursery type, four wind-direction columns), which were created through the label method. The data was then split randomly, using a fixed random seed so that results would be reproducible between different data splits, between a 80–20 train–test split (1870 records in the train set and 468 in the test set).

C. Exploratory Analysis

Plotting on box and whisker plots revealed a few high magnitude values for these two factors, pesticide dosage and trash quantity, but these values had a high degree of plausibility and were not considered erroneous entries, but rather as true variation in the study field. After testing for generally low-to-moderate correlation between the inputs of each agronomic (Fig. 1), this suggests very few instances of multicollinearity between them, and moderate positive correlation between temperature readings taken at similar time of year for adjacent growth periods (as weather should be changing gradually and continuously), except for the output of growth 2, which had moderate-positive correlation with both growth 1 and growth 3.

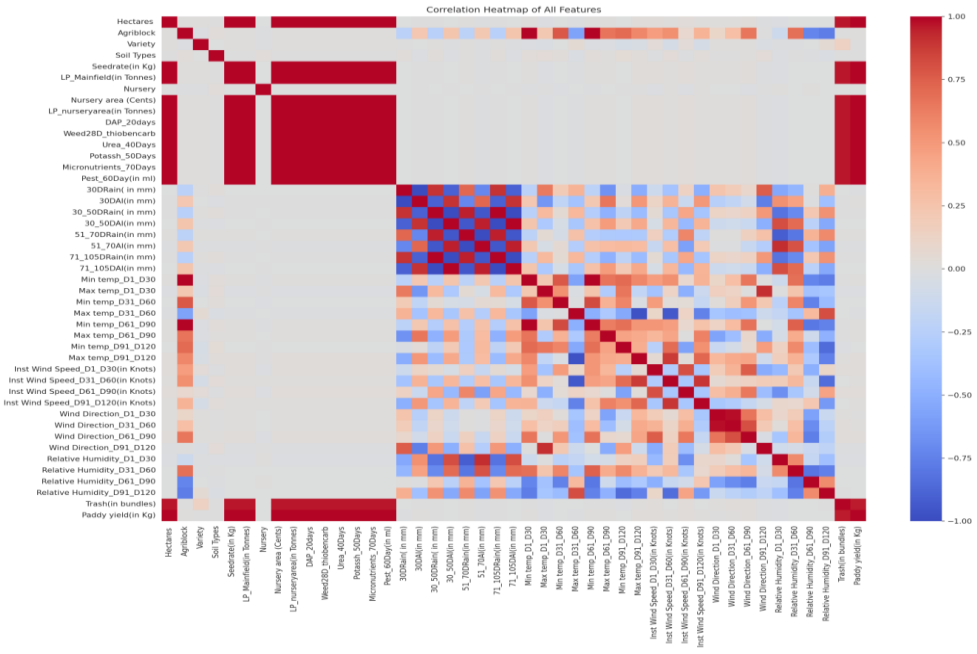

Fig. 1. Correlation heatmap of dataset features.

D. Feature Selection Strategy

For modeling all 44 independent features were used as input for the regressor Random Forest, which in turn calculates importance value for every variable using the reduction of prediction error resulting from its absence from the different trees which make up the regression model. As opposed to a pre-judging of factors, this approach proved itself to be a model-driven, where the most important agronomic and climatic factors were determined by the observed data, and the resulting rankings are presented in the Section IX.

VI. SYSTEM ARCHITECTURE

The proposed system consists of four-layer system. The raw spreadsheet containing cultivation information (data) is ingested by the Data Layer. Pre-processing Layer removes duplicate, imputes missing values, makes categorical encoding and applies feature scaling. In the Model layer, it trains the 6 candidate regressors and tunes the selected model (trained in Model layer). The Evaluation and Output Layer calculates accuracy metrics, generates visualisations for comparison and writes the final trained model to save for later use. This separation of concerns allows each stage to be tested in isolation, and enables the inclusion of new algorithms, features, or data sources without disruption to the pipeline's other components.

The processing sequence is: (1) load the raw data; (2) apply initial data exploration and integrity checks; (3) remove duplicate rows; (4) fill in missing values as a precaution; (5) apply label encoding for categorical data; (6) separate the target variable; (7) conduct preliminary exploratory data analysis; (8) standard scale all data inputs; (9) split the data into train, test, and validate sets; (10) train six regression models; (11) measure each model using MAE, MSE, RMSE and R^2; (12) extract feature importances from the Random Forest model; (13) perform five-fold cross validation; (14) tune the Random Forest model with GridSearchCV; (15) persist the tuned model for future predictions.

VII. Utilizing Algorithms and Models.

A. Linear Regression

Estimates a linear combination of the features provided input to the target value, and serves as a reference for other estimators, such as non-linear estimators, to be compared to.

B. Decision Tree Regressor

A recursive strategy into partitioning the feature space into regions that minimise prediction error; the problems of fitting non-linear interactions in the features while at the same time being liable to overfitting if the search is left unconstrained.

C. Random Forest Regressor

Results in prediction of multiple decision trees trained on a bootstrapped subset of the data and a random subset of the features, and averages their prediction. In this study after tuning, the best model was found to be Random Forest whose variance reduced by using this variance-reducing technique [1] for this particular data.

D. Support Vector Regression (SVR)

Construct a function that fits a given error bound and has a limited model size [4]. It gives unreliable results when the kernel and regularization parameters as well as feature scaling parameters are not carefully tuned, as was a drawback in this data-set.

E. K-Nearest Neighbours (KNN)

Estimates a target by using the average of the k nearest neighbors, in the feature space, based on the Euclidean distance, used k = 5 in this study.

F. Gradient Boosting Regressor

Constructs a shallow tree with additive regression to reduce the error remaining from the ensemble constructed thus far [2]. The difficulty was that XGBoost could not be installed in the available execution environment and Gradient Boosting is a well-known technique with the same underlying boosting principle [5] but is easier to use as it is already available.

VIII. IMPLEMENTATION

A. Tools and Environment

It is implemented in Python 3 within a Jupyter Notebook and involves different libraries such as pandas and NumPy for data manipulation, matplotlib and seaborn for data visualization, scikit-learn for model training, evaluation, and tuning [3] and joblib for model persistence. The whole process is implemented on standard desktop systems without special computing facilities, so as to make practical use of the process by the extension and student researchers.

B. Model Training and Testing

Each model's performance was tested with the same held out test partition, which comprised 20% of the dataset, with all models trained on the same training partition (80%).To fairly compare model performance, each model was trained on the same 80% partition and tested on the same 20% held out partition. After obtaining one of the most promising candidates, the model performances of the new trained model were searched using GridSearchCV over the number of estimators (100, 150), maximum tree depth (5, 10) and minimum samples per split (2, 5) under three-fold cross validation. The optimal setting found by the search yielded 150 estimators, 5 maximum depth and 5 minimum split size for an R^2 cross-validated equal to 0.99.

C. Validation Procedure

Five-fold cross validation was also applied to the whole standardized set to ensure that the reported accuracy did not appear to be sensitive to the division of the train and test sets. Initially all folds were folded out for validation, with the remaining folds used for training the model; a mean R^2 of 0.99 with a standard deviation of 0.00 across folds was obtained – indicative of the model being well split independent and stable.

Hyperparameter	Search Space	Selected Value

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 511-523 March 2026-Sept 2026 www.csjournals.com

n_estimators	100, 150	150
max_depth	5, 10	5
min_samples_split	2, 5	5
Best CV R^2	—	0.99

TABLE II. GRIDSEARCHCV SEARCH SPACE AND SELECTED HYPERPARAMETERS

IX. RESULT AND DICUSSION

A. Comparative Model Performance

We evaluated six different algorithms on the 468-record test set, and the results are in Table III, showing MAE, RMSE, and R^2. The optimized Random Forest model performed best, with the lowest error and highest explanatory power ($R^2 = 0.99$, MAE ≈ 655 kg, RMSE ≈ 910 kg). Gradient Boosting and the standard Random Forest weren't far behind, both achieving an R^2 over 0.98. Decision Tree, KNN, and Linear Regression models were slightly less accurate, but still reasonably good. The SVR model was a clear outlier, showing a negative R^2. This suggests its predictions were worse than just guessing the average yield, likely because it struggled with the dataset's high dimensionality and varied feature scales. We saw similar issues with SVR in Section II.

Model	**R^2**	**MAE (kg)**	**RMSE (kg)**
Tuned Random Forest	0.99	655	910
Gradient Boosting	0.99	679	947
Random Forest	0.99	696	969
Decision Tree	0.99	702	980
KNN	0.99	714	996
Linear Regression	0.99	769	1,026
SVR	-0.01	7,799	9,293

TABLE III. PERFORMANCE COMPARISON OF SIX REGRESSION MODELS ON THE TEST SET

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 511-523 March 2026-Sept 2026 www.csjournals.com

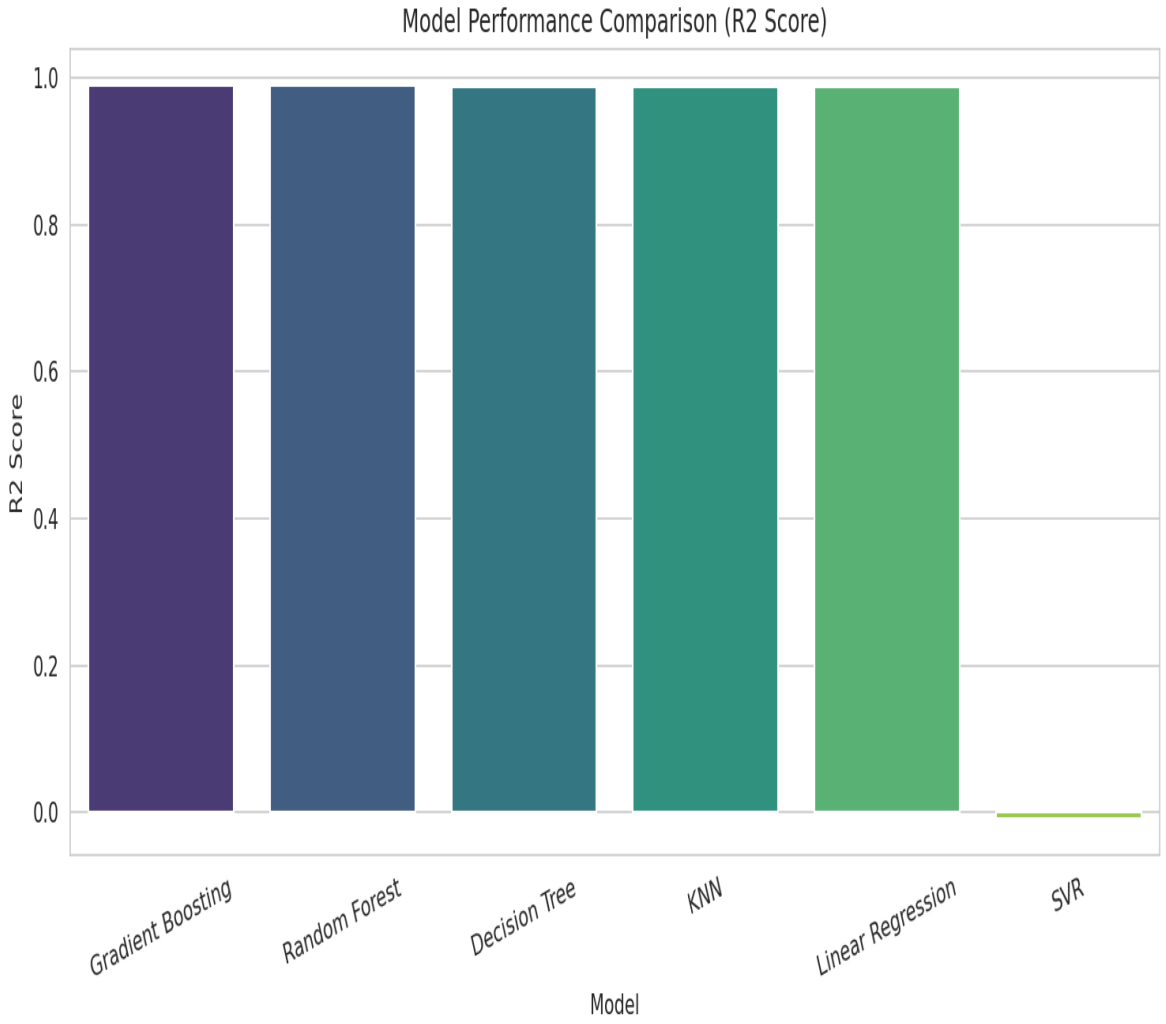

Fig. 2. R^2 comparison across the six regression models.

B. Feature Importance

Feature importance scores from the Random Forest model (Fig. 3) indicate that five most impactful factors are seed rate, DAP application at 20th day, micronutrient application at 70th day, nursery area, and land-preparation work for the nursery area. Together, they explain a significant portion of model's power. Static attributes such as soil type, nursery type, and paddy variety were least of the concern which indicates that in such range of dataset conditions input management is more important than fixed categorical classification.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 511-523 March 2026-Sept 2026 www.csjournals.com

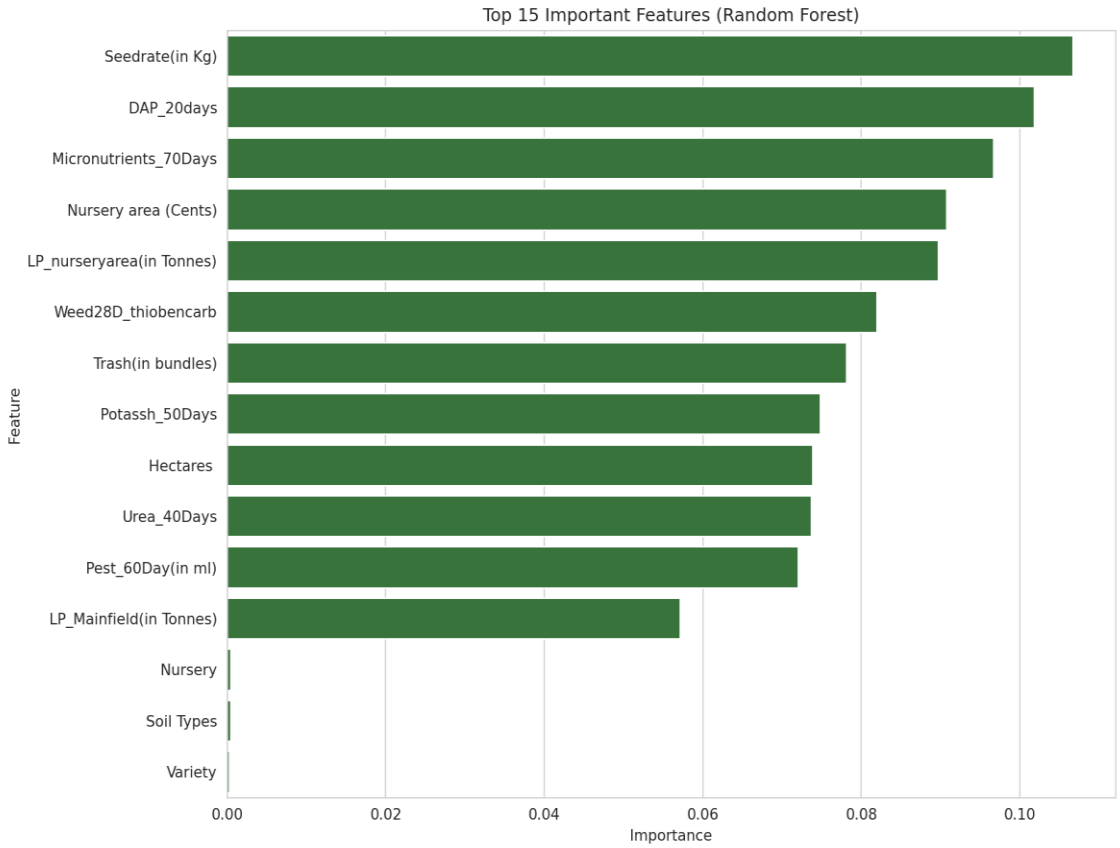

Fig. 3. Top-ranked feature importances from the Random Forest model.

C. Residual Behaviour and Robustness

Fig. 4 presents the graph of predicted yield against actual yield for the tuned Random Forest model; the points are grouped closely along the diagonal throughout the 5, 410-38, 814 kg range, meaning that the model does not have a tendency to consistently over-predict or under-predict at either end of the spectrum. The average absolute error of 655 kg is just about 2.9 % of the mean yield of the dataset, and also there was no significant heteroscedasticity observed, thus prediction error does not increase disproportionately at higher levels of yield. Five-fold cross-validation (mean R^2=0.99, STD=0.00) also demonstrated that the model not only fits well on one subset but maintains accuracy across different subsets.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 511-523 March 2026-Sept 2026 www.csjournals.com

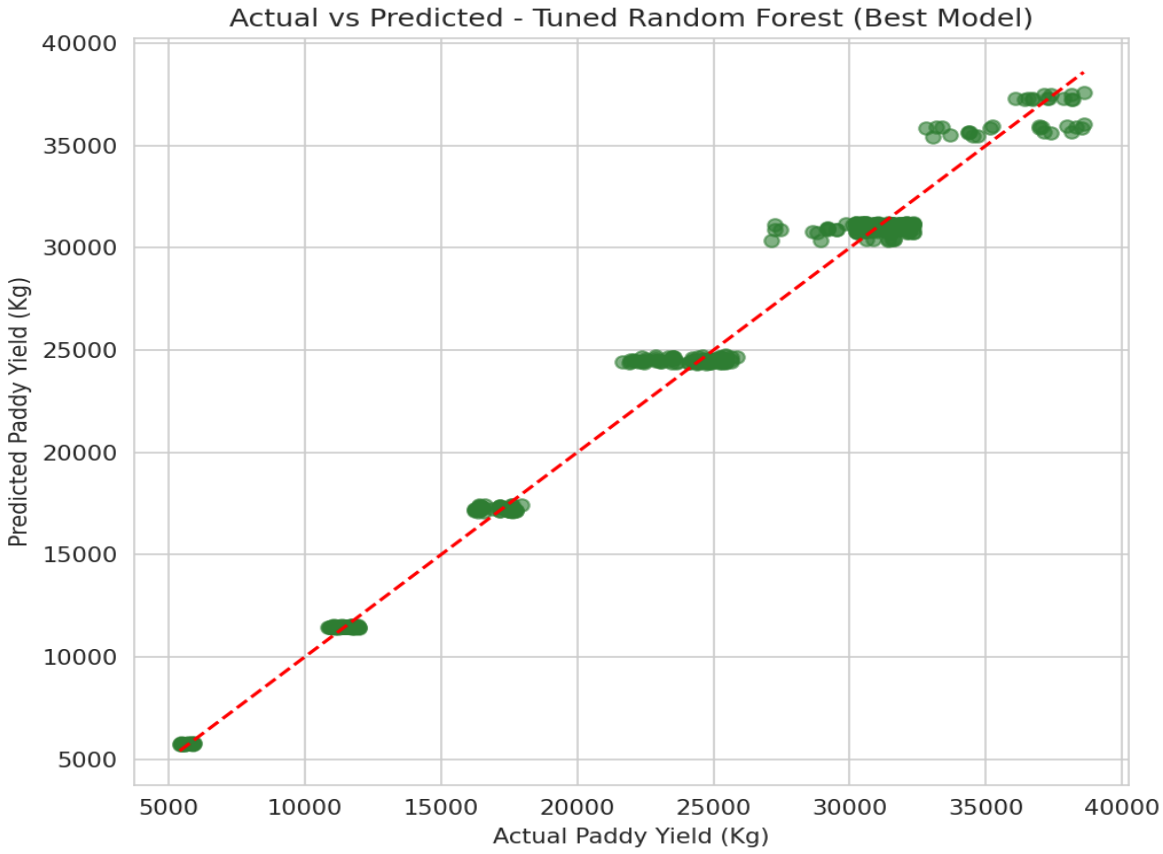

Fig. 4. Actual versus predicted yield for the tuned Random Forest model.

D. Discussion

These results confirm the main conclusion of the literature on the topic: ensemble tree-based regressors are better capable of modeling the non-linear and interacting structure of the agronomic and climatic data than linear or kernel-based methods. The very poor performance of SVR even after feature standardization shows that algorithm selection should be data-dependent and not based on a single technique by default. On a practical level, the fact that seed rate, DAP dosage, and the timing of the micronutrient application are far more important than static categorical variables, means that the extension advice that targets managing precisely and timing the inputs carefully will likely deliver much more benefit than the advice that is only based on soil-type or variety classification, although the climatic conditions during the growth period remain a very important setting factor.

X. CONCLUSION

This paper put forward and tested a machine learning model for predicting paddy yield based on a field-level dataset that merges agronomic management practices with climatic data gathered stage-wise. Of the six regression models that were compared using the same protocol, the Random Forest model that was fine-tuned by GridSearchCV showed the best and the most consistent results with an R^2 of 0.99 and an RMSE of around 910 kg, as verified by five-fold cross-validation. Feature-importance analysis brought out seed rate, DAP dosage,

micronutrient timing, and nursery area as the major factors influencing yield, far outweighing the static categorical attributes such as soil type and variety. Therefore, the model proves that a data-driven regression approach can greatly exceed the accuracy of a guess based on experience, and provides a useful foundation for decision-support tools aimed at assisting farmers and agricultural planners in more efficiently allocating inputs and achieving better rice yields.

XI. FUTURE SCOPE

1. Expanding the current six agricultural blocks dataset with other states, different soil types, and various climate zones to increase the model's applicability level for the entire country.

2. Use short-term weather forecasts and IoT-enabled soil-moisture, and soil nutrient sensors to facilitate the generation of forward-looking, in-season predictions.

3. Use sequence-aware deep learning approaches such as LSTMs to model the step-wise climatic data time series more effectively.

4. Use satellite-based vegetation indices (e.g., NDVI) to add a direct crop-health signal to the current tabular feature set.

5. Make the refined model accessible via a web or mobile platform so that the end users, viz. farmers and field officers, can get instant yield estimates and input-optimization recommendations.

6. Compare XGBoost in a plug-and-play manner with the Gradient Boosting substitute used here, once the library issues have been resolved; and extend the current framework to other major food crops.

. REFERENCES

[1] L. Breiman, "Random forests, " Machine Learning, vol. 45, no. 1, pp. 5, 32, 2001.

[2] J. H. Friedman, "Greedy function approximation: A gradient boosting machine, " The Annals of Statistics, vol. 29, no. 5, pp. 1189, 1232, 2001.

[3] F. Pedregosa et al., "Scikit-learn: Machine learning in Python, " Journal of Machine Learning Research, vol. 12, pp. 2825, 2830, 2011.

[4] C. Cortes and V. Vapnik, "Support-vector networks, " Machine Learning, vol. 20, no. 3, pp. 273, 297, 1995.

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system, " in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785, 794.

[6] W. McKinney, "Data structures for statistical computing in Python, " in Proc. 9th Python in Science Conf., 2010, pp. 51, 56.

[7] J. D. Hunter, "Matplotlib: A 2D graphics environment, " Computing in Science & Engineering, vol. 9, no. 3, pp. 90, 95, 2007.

[8] M. L. Waskom, "seaborn: Statistical data visualization, " Journal of Open Source Software, vol. 6, no. 60, p. 3021, 2021.

[9] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review, " Sensors, vol. 18, no. 8, p. 2674, 2018.

[10] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review, " Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020.

[11] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.

[12] Government of India, Ministry of Agriculture & Farmers Welfare, Annual Report on Rice Production Statistics.

[13] Food and Agriculture Organization of the United Nations, Rice Market Monitor Reports.