Rice crop yield prediction using hierarchical time series forecasting with ensemble techniques in Telangana region

P. Sowmya[1*] , A.V. Krishna Prasad[2]

(1. Department of Computer Science, Telangana Mahila Viswa Vidyalayam, Koti, Hyderabad, Telangana 500095, India; 2. Department of Computer Science, Maturi Venkata Subba Rao Engineering College, Nadargul Main Rd, Hyderabad, Telangana 501510, India)

Abstract: Agriculture is the main source of income especially rice production for the great majority of people. In agriculture, accurate rice crop yield forecasting is crucial for making various crop production-related policy decisions that will assure the supply of food. Taking into account, many works concentrating on machine learning and ensemble based techniques; however, it has several limitations such as increased complexity, overfitting and under fitting, struggle to generalize effectively. Therefore, to overcome the limitations this research proposed a novel rice crop yield prediction using hierarchical time series forecasting with ensemble techniques in Telangana region. In the initial stage, the collected dataset is pre-processed to improve the accuracy of the prediction in which the missing values in imputed by k-Nearest Neighbors (k-NN) imputation method and normalization used to normalize the input data. Then, this pre-processed input data is start by organizing the crop yield data into a hierarchical structure by using grid based AutoRegressive Integrated Moving Average (ARIMA) model. Moreover, this research utilizing Extreme Gradient (XG) Boosting as a meta-model and train the meta-model using the new dataset. As a result, by combining the ARIMA-based Hierarchical time series forecasting model with a stacking ensemble provides a higher accuracy.

Keywords: crop yield prediction, real time data, ensemble approach, artificial intelligence approaches.

Citation: Sowmya, P., and A. V. Krishna Prasad. 2025. Rice crop yield prediction using hierarchical time series forecasting with ensemble techniques in Telangana region. Agricultural Engineering International: CIGR Journal, 27(2):285-298.

1 Introduction

Agriculture is the backbone of most countries around the world, supplying food and textiles. The world's population is quickly expanding, and this must have an impact on food production. As a result, precision agriculture and conventional farming must be integrated (Rajak et al., 2017; Yesugade et al., 2018). The worldwide production of food has to treble by 2050 to fulfil the needs of the world's constantly expanding inhabitants (Do Land, 2009; Tilman et al., 2011). To achieve this goal, however, the current yield growth rates for main grains cultivated worldwide are insufficient (Ray et al., 2013). The agricultural sector presents significant challenges due to changes in the environment, including the effects of global warming and variations in the climate (Bhadouria et al., 2019). Food insecurity might increase globally because of this decreasing agricultural productivity (Deryng et al., 2014). Crop production forecasting is one of the most challenging problems in precision agriculture. With the use of crop production predictions, pertinent

Received date: 2024-03-06 **Accepted date:** 2024-10-03

***Corresponding author: P. Sowmya,** Department of Computer Science, Telangana Mahila Viswa Vidyalayam, Koti, Hyderabad, Telangana 500095, India. Email: p.sowmya2105@gmail.com.

authorities can decide the best way to guarantee food security. Crop yield is significantly influenced by climate and soil conditions, genotype, and management strategies (Singh et al., 2014). Weather-related losses account for around 30% of annual productivity worldwide (Attri and Rathore, 2003). Because of this, there is a great need for models that offer accurate production forecasts before harvest, so that producers, policymakers, and the government may make advance plans (Shahhosseini et al., 2021; Palanivel and Surianarayanan, 2019).

Over half of the world's 7.5 billion people are fed by rice (Singha et al., 2019), with Asia producing 90% of the rice consumed globally and 87% of the harvested land for cultivated rice (Zhang et al., 2020; Zhang et al., 2017). Almost 20% of the rice produced worldwide originates in India, which is the second-biggest manufacturer in the world after China (FAOSTAT, 2022). But in the most susceptible area of the world—the tropics—monsoon fluctuation brought on by climate change has a significant impact on rice yield in India (Gupta and Mishra, 2019; Soora et al., 2013). It is challenging to increase food production in response to the pressures of climate change and fulfil the growing demand of the population (Zabel et al., 2021). Therefore, accurate, reliable, and accurate rice production forecasts is essential for advertising scheduling, health, and local, national, and worldwide food security in India (Feng et al., 2021). Crop output may be predicted using a variety of methods, including predictive machine learning models, field experiments, and crop development models. For massive amounts forecasts, numerical machine learning models and simulations of crop development are more reliable and practical than tedious and difficult field surveys. These simulations, however, require enough field data to alter certain parameters, limiting their efficacy at regional scales. Machine learning algorithms can eliminate crop-specific parameter dependence and term nonlinear correlations between input data and crop output (Paudel et al., 2021; Cai et al., 2019; Ma et al., 2021). Unfortunately, there hasn't been much research done to date on developing a model to predict rice crop production for the Telugana region using a hierarchical time series forecasting model and ensemble technique. The main influence of this study is as follows:

1) In the initial stage, the collected data's cleaned to improve the accuracy of the prediction in which the missing values in imputed by k-Nearest Neighbors (k-NN) imputation method and normalization used to normalize the input data.

2) Then, we are using grid search method used to tune the AutoRegressive Integrated Moving Average (ARIMA) hyper-parameters for each time series in the hierarchy.

3) Then, combine the base level forecasts and actual crop yield data from the test set to create the new dataset for the meta-model which has base-level forecasts. In our work, XGBoosting is used as a meta-model and train the meta-model using the new data's.

The design of this investigation's paper is as follows: After reviewing previous crop yield predictions in Section 2, Section 3 describes the proposed crop yield projection. Section 4 presents the outcomes of the proposed approach's execution, and Section 5 offers the conclusion.

This section addresses some important studies on crop yield prediction. The researcher makes use of a number of strategies. Here, we go through a few of the important pieces. Based on weekly weather indices, using artificial intelligence algorithms, Das et al. (2018) developed a variety of crop growth estimation algorithms for the fourteen regions on the west coast. A random forest approach, as proposed by Kamath et al. (2021), offered a quick review of agricultural yield forecasts. Given the large volume of data needed for crop output predictions, data mining techniques are a perfect fit. Li et al. (2021) have presented a random forest model-based variable yield estimation scheme. It was feasible to forecast crop yields and have a better knowledge of how yields respond to changing climatic circumstances by using the recommended strategy. According to Iniyan and Jebakumar (2022), the newly developed composite

linear crop forecasting algorithm outperformed several supervised neural networks and advanced ensemble learning techniques. In terms of expected yield, the highly developed ensemble regression crop prediction model fared better than several supervised machine learning and advanced ensemble learning techniques. In order to evaluate how effectively the fundamental approach works with respect to certain metrics of performance under varying weather and soil characteristics, Oikonomidis et al. (2022) suggested employing a deep learning model. Machine learning (ML) algorithms were all examined, in addition to the XGBoost ML method. Principle Component Analysis (PCA) was suggested by Nain et al. (2021) as a solution to the divergence issue. The accuracy of yield forecasts is improved by using PCs from weather observations as predictor factors. A multivariate technique called discriminate analysis involves classifying and allocating new items to already defined groups. The harvest rate may also be predicted using a weather parameter forecast and an exploratory achieve regressor.

Ajithkumar et al. (2021) developed a several-weather-based mathematical framework that used aggregate weather factors and principal component extraction to anticipate the productivity of two distinct rice cultivars. The combined impact of climatic factors may be represented using the Principal Component Regression (PCR) and composite weather variables (CWV) models, respectively. The amount of rice crop predicted by the goodness of fit of the individual programmers was determined using the t-test. The calculated value of t was found to be smaller than the t-critical value in both scenarios. Consequently, it was found that the projected yield and the actual rate of return were quite close. In order to predict agricultural production, Kundu et al. (2022) used new machine learning and artificial intelligence techniques to forecast yields of crops utilising a variety of data factors, including the climate, the earth, producers, cultivation, and chemical. Chandraprabha and Dhanraj (2023) proposed Model Agnostic Meta Learning (MAML), a stacking-based collaborative learning system, for rice yield predictions in a given soil. Data on crop productivity and soil nutrients were utilised as inputs in the suggested technique. Apat et al. (2022) presented an assessment assistance framework and an Internet of things with heterogeneous ensemble learning environment (IoT-HELE) -based smart-farming prognosis and intelligent agricultural analytics model that employ state-of-the-art machine learning and deep learning approaches to accurately anticipate crop yield. In this approach, ensemble voting leads to a more lucrative, sustainable, and effective agricultural business.

As a result, from the above analysis ensemble methods are not concentrating the prediction in various levels especially in the lower levels which lead to decrease the prediction accuracy, the existing methods are not understanding the yield fluctuations from the patterns and trends. Also, these methods are missing the uncertainty levels in the yield predictions.

2 Materials and method

India is an economically developing nation whose economy is mostly based on agriculture. The diversity of agricultural resources including soil, fertiliser, and environmental factors, such as crop conditions. It is crucial to use these resources properly for optimum output. Because of practises based on human experience, particularly given the erratic environment, crop yields are low. Here, many works concentrating ensemble techniques to improve the accuracy, still they are facing problems such as increased complexity, right combination of models requires careful tuning to ovoid overfitting or under-fitting and struggle to generalize effectively when faced with extreme or unseen data patterns, especially if the training data is not representative of future scenarios. In order to get over the aforementioned restrictions, this study suggests a novel crop production forecast, which is displayed in Figure 1.

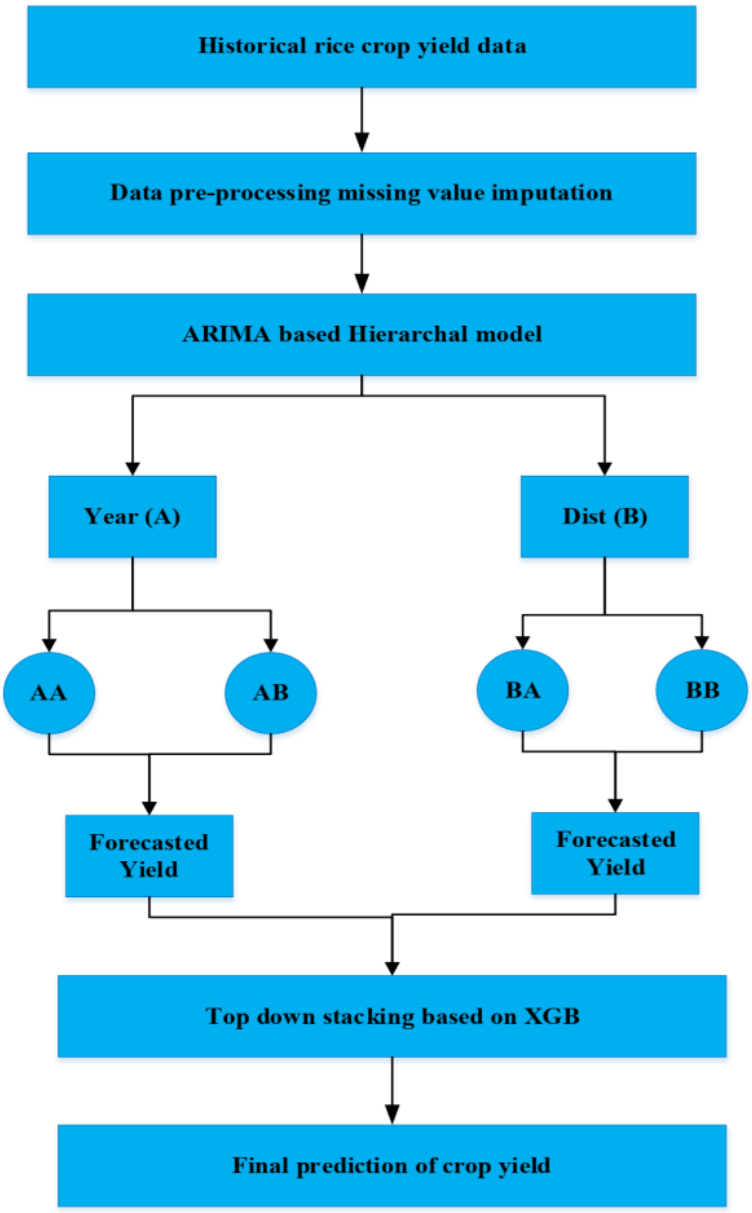

Figure 1 Construction of the proposed approach

2.1 Pre-processing

In the initial stage, the collected data's are cleaned to improve the accuracy of the prediction in which the missing values in imputed by k-NN imputation method and normalization used to normalize the input data.

K-nearest neighbors (KNN) is frequently used for its ease of use and track record for performance with several imputation of missing value issues. Two sorts of data may be predicted using the higher accuracy of the KNN imputation method: discrete information (mode value) and continuous information (mean value). In this study, we processed continuous information. Creating an estimation model for every data criterion with values that are missing is not necessary when imputing using KNN.

Calculate K, the amount of closest observations that were used. Using the Equation 1 for calculating the Euclidean distance, determine the distance between data with missing values in jth and observations without missing values on the variable.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (1)$$

Where, $d(x_i, x_j)$ is the distance from i to the midpoint of cluster j; x_i is the preparation information, x_j is the analysis information; n is the amount of characteristics; k is the characteristic; x_{ik} is the ith data on the kth characteristic, x_{jk} is the jth

data on the kth characteristic.

Using the least range estimate as a guide, find the smallest k measurements. When a measurement has an insufficient value, the value of j in the smallest k occurrences will be utilized in the estimation procedure. Fourth, figure us the amount of weight each of the k shortest occurrences carries. The score that is highest will go to the nearest observation. Fifth, use Equation 2 to determine the mean value in the minimum k occurrences that do not contain a missing value.

$$X_j = \frac{1}{k}\sum_{k=1}^{k} v_{kj} \quad (2)$$

Where, k is the nearest observation that was utilised, v_{kj} is the value of the entire data on the missing parameter value, and X_j is the weighted mean. Sixth, use the mean value acquired at stage 5 to complete the estimation procedure for the values that are lacking for observations that have missing values.

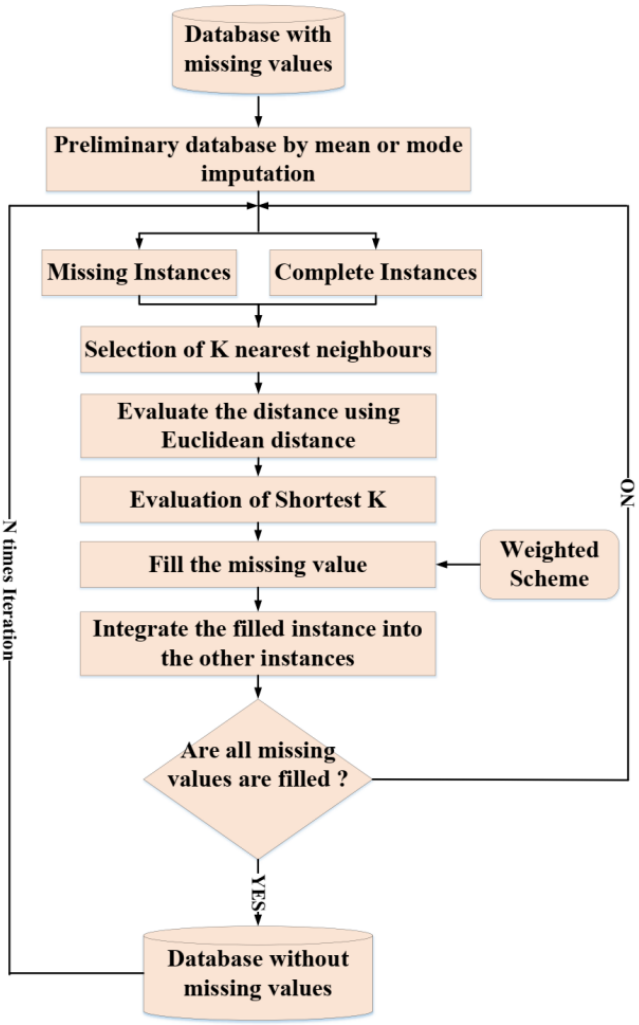

Figure 2 Architecture of KNN imputation

The following contains the normalisation approach details: Standard scaler normalisation is the term for the method that uses variables such as average and standard deviation to provide normalised values or an assortment derived from the original unprocessed information. Therefore, using the conventional scaler value and the following formulas, unstructured information may be normalised:

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

Where, z is the normalised significance, x is the characteristic's initial significance, σ is the feature's average deviation, and μ is the mean value.

Assume for the purposes of this approach that there are five rows—X, Y, Z, U, and V—each containing a distinct variable or category that is represented by the character "n." Therefore, the normalised ones may be calculated in each row above using the conventional scaler approach. All values for

a row have the value zero if, like in the case of the hypothetical row with all similar values, the row's standard deviation equals zero.

Algorithm 1: Standard Scaler Normalization

function standard scaler_normalization (data):

For each feature in data mean = calculate_mean_feature // Compute the mean of the feature.

std.dev = calculate_standard_deviation (feature) // Calculate the standard deviation of the feature

For each data_point in data: data_point[feature] = (data_point[feature]-mean) / std.dev

Return data

Then, this pre-processed input data is start by organizing the crop yield data into a hierarchical structure based on the different levels of aggregation such as year and district of the Telangana region.

2.2 Arima models

In our work we are using ARIMA as a hierarchical time series forecasting model and train the historical rice crop yield data at each level of the hierarchy to make individual forecasts and grid search method used to tune the ARIMA hyper-parameters for each period sequences in the ladder. We train the ARIMA model on the training data for each level and validate their performance on the test data.

The range and median of ARIMA models must remain stable across periods as they are designed to analyse stagnant dynamics. Utilising the increased Dickey-Fuller (ADF) assessment, we can determine if the operation is stable or not (yet). Three essential components are used by ARIMA to describe historical a sequence:

- Autoregressive terms (AR), which simulate historical process data.

- Integrated terms (I) that simulate the variations required to bring the system to a stable state.

- The amount of noise around the process is controlled using the moving average (MA), which is a statistical measure.

The autoregressive phase (p), differencing phase (d), and moving average order (q) are the hyperparameter that for ARIMA models. For every parametric, create an array of potential values. In particular, the sentences AR provide an interpretation of the sequence according to its historical data points:

$$X_{st} = \emptyset_1 X_{st-1} + \emptyset_2 X_{st-2} + \cdots + \emptyset_p X_{st-sp} = \sum_{j=1}^{sp} \emptyset_j X_{st-j} \quad (4)$$

Where, X represents the predicted value of the target time series at a particular point in time, X_{st} denotes the value of time series at time step st, and the sequence ($\emptyset_j$) j denotes the auto-regressive parameters. Rather of emphasising historical data, the MA constituent displays a rolling average of prior error variables in the model of regression (innovation processes):

$$X_{st} = \emptyset_1 \varepsilon_{st-1} + \emptyset_2 \varepsilon_{st-2} + \cdots + \emptyset s_p \varepsilon_{t-sp} = \sum_{j=1}^{sq} \emptyset_j \varepsilon_{st-j} \quad (5)$$

Where, st is the amount of average motion leaps used to anticipate the present state of the sequence, sq is the amount of methods of innovation that have occurred before (ε_{st-j}) j, and the series of $(\emptyset_j)$ j indicates the MA parameters. Further to all of this, there is also the merged component, or ARIMA, which is determined by an integer d that denotes the differentiation sequence:

$$X_{st}^* = X_{st} - X_{st-1} - \cdots - X_{st-sd} \quad (6)$$

In most cases, adopting $d = 1$ is sufficient. The ARMA class of operations, for which $d = 0$, is a noteworthy group of processes. The Backward operator $BX_{st} = X_{st-1}$ is introduced, allowing for the formal formulation of an ARIMA model:

$$\left(I - \sum_{j=1}^{sp} \emptyset_j B^j\right)(I - B)^d X_t = \left(I + \sum_{j=1}^{sq} \emptyset_j B^j\right)\epsilon_{st} \quad (7)$$

Additionally, Box and Jenkins expanded the ARIMA approach to identify variations in the seasons in a time sequence. As a result, we refer to the SARIMA model (the additional S standing for "Seasonal") and write,

$$X_{st} = SARIMA(a, b, c)(A, B, C)_m \quad (8)$$

The amount of incorporated, portable, and auto-regressive periodic components is indicated by the variables A, B, and C, respectively. The periodic

phenomenon's annual periodicity is determined by m. Then, combine the base level forecasts and actual crop yield data from the test set to create the new dataset for the meta-model which has base-level forecasts as the input features and actual crop yield as the target variable.

For top-down stacking ensemble, in our work we are using XGBoosting as a meta-model and train the meta-model using the arrived dataset from the ARIMA model.

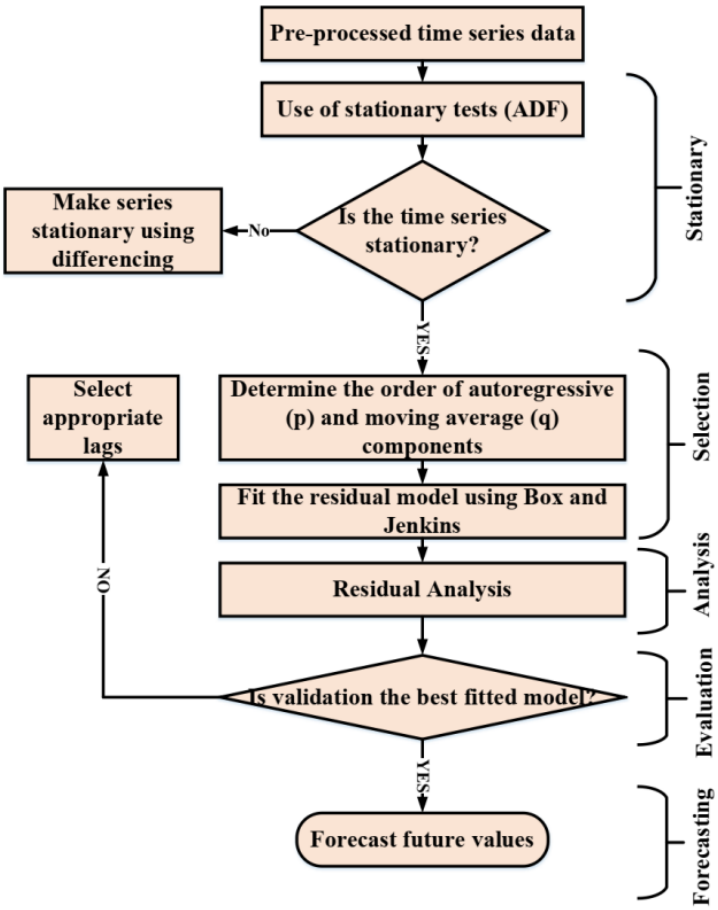

Figure 3 Architecture of ARIMA models

2.3 Top-down stacking ensemble

This proposed system used his trained meta-model to combine the base-level forecasts from ARIMA with the test set data from each level of the hierarchy and generate aggregated forecasts which consider the individual forecasts from ARIMA and other relevant information from the test set .

Considered of the gradient enhancement device (gbm) executions, Xgboost, is one of the most effective strategies used in learning with supervision. Because of its rapid execution speed, Xgboost is the most recommended. Every tree in the cumulative conditioning method of boosting is constructed by taking insights from the leftover δ of the preceding tree. $\hat{y}_i^k = \hat{y}_i^{t-1} + f_t(X_i)$ is the forecast of the k-th repetition. XGboost reduces the anticipated loss and optimises the algorithm at every phase. The stacked summing of trees produces the final forecast output $\hat{y}_l$ in the way described below:

$$\hat{y}_l = \sum_{t=1}^{T} f_t(X_i), f_t \in \mathcal{F} \quad (9)$$

Where, $\mathcal{F}$ is the planetary of purposes comprising all reversion trees; T represents the amount of trees. To acquire purpose f_t of each tree, XGBoost creates an independent task with regularization:

$$\mathcal{L}(\Phi) = \sum_i l(y_i, \hat{y}_l) + \sum_t \Omega(f_t) \quad (10)$$

Where, Φ is all learnable constraints in XGBoost; $l(y_i, \hat{y}_l)$ is the defeat task demonstrating the inaccuracy between the forecast crop yield $\hat{y}_l$ and the real yield y_i, the lesser the l is, the improved the performance of the procedure; $\boldsymbol{\Omega}(f_s)$ is the regularization period to correct the typical difficulty and avert over-fitting. When XGBoost uses the

square defeat task to amount error, the second derived Taylor growth of the defeat task can support the archetypal to enhance the objective rapidly. The second derived Taylor growth of the defeat task after $t-th$ repetition is given as follows:

$$\mathcal{L}(\Phi)^{(s)} = \sum_{i=1}^{n}\left[l\left(y_i^{(s)}, \hat{y}_i^{(s-1)}\right) + g_i f_s(x_i) + \frac{1}{2}h_i f_s^2(x_i)\right] + \Omega(f_s) \quad (11)$$

Where, g_i and h_i are the impairment function's initial and subsequent derivatives, respectively. It may be discovered that every information point's initial and subsequent derivatives are the only things that the coefficient of loss depends on. Optimising the XGBoost algorithm variables, booster variables, and learning variables is a crucial phase in the learning algorithm that assists in estimating rice crop production.

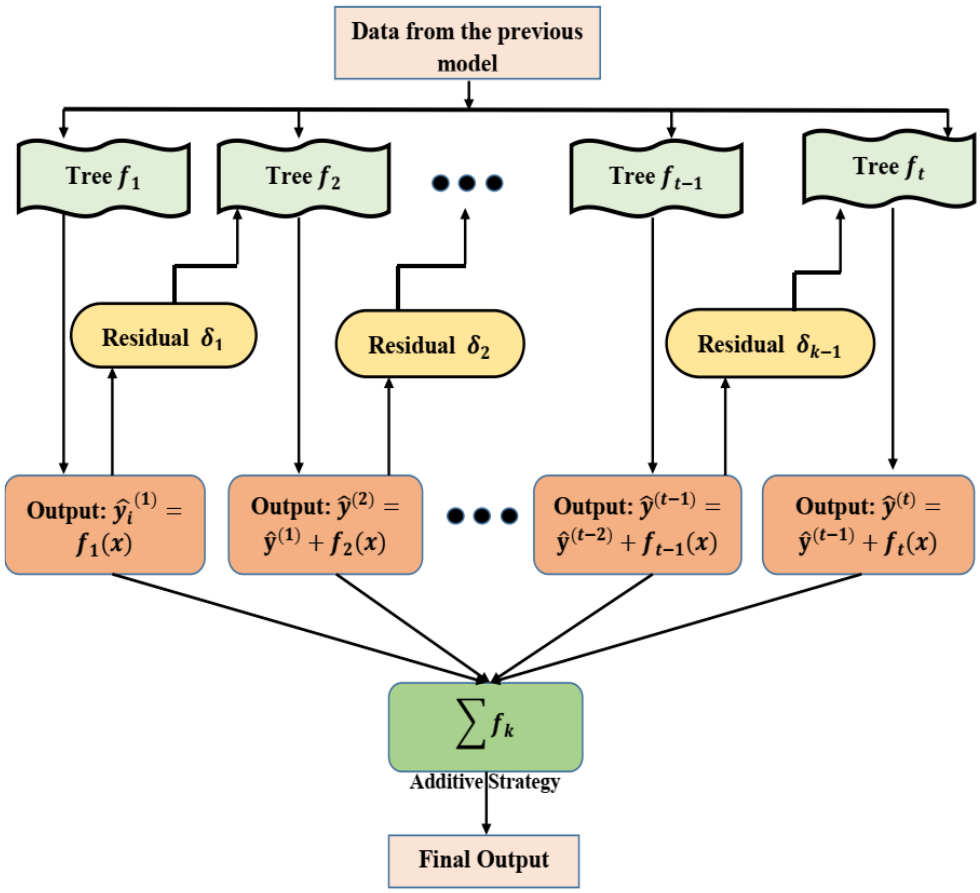

Figure 4 Architecture of XG boost

As a result, by combining the ARIMA-based Hierarchical time series forecasting model with a stacking ensemble, we can leverage the individual strengths of both approaches and potentially achieve more accurate and robust crop yield predictions, making it well-suited for the complex and nuanced nature of agricultural forecasting.

3 Results and discussion

This part discusses our offered technique's efficiency as well as the outcomes of its construction. Additionally, the baseline approach's comparative findings are also covered.

Tool: PYTHON 3

OS : Windows 7 (64-bit)

Processor: Intel Premium

RAM : 8GB RAM

3.1 Materials and Methods

This dataset encompasses a comprehensive collection of diverse variables spanning a range of parameters across various districts and one states. It contains essential information crucial for understanding agricultural dynamics, climatic patterns, and soil conditions within specific districts in a country or region.

Overview of the Dataset:

District Code: Unique identifier for each district within the region or country.

Year (1966-2020): Temporal information indicating the year of data collection, spanning several decades from 1966 to 2020, providing a long-term perspective on trends and changes.

State Code: Identifies the state to which the district belongs, allowing for regional comparisons and analysis.

District Name: Names of the districts under study, providing geographic context to the data.

Rice Irrigated Area: Quantifies the area under rice cultivation that is irrigated, indicating the extent of agricultural practices related to rice farming and irrigation methods used.

Total Consumption: Measures the overall consumption within the district, potentially encompassing various resources or goods related to agriculture, living standards, or other socioeconomic indicators.

Annual Rainfall: Records the total amount of precipitation received in the district over the course of a year, providing critical information about the local climate and its impact on agricultural practices.

Precipitation: Specific data on precipitation levels within the district, which may offer insights into short-term weather patterns and their effects on agriculture.

Maximum and Minimum Temperature: Records the highest and lowest temperatures experienced within the district, contributing to an understanding of the local climate's temperature range and its implications for crop growth and farming practices.

Wind Speed and Direction: Details regarding wind characteristics in the district, which can influence various aspects of agriculture, including pollination, pest control, and crop damage.

Surface Soil Wetness: Measures the moisture content of the surface soil, an essential factor impacting crop growth and agricultural productivity.

Profile Soil Moisture: Reflects the moisture content deeper within the soil profile, providing insights into long-term soil moisture conditions and potential water availability for plants.

Root Soil Wetness: Indicates the moisture levels in the soil at the root level of plants, influencing plant health, growth, and overall crop yield.

Rice Yield: Quantifies the amount of rice harvested per unit area, serving as a primary indicator of agricultural productivity and success.

3.2 Experimental Results

3.2.1 ARIMA forecasting results

Analysing the historical trends in rice yield across various districts or regions has provided a foundation for building the ARIMA model. The dataset, encompassing yearly or seasonal observations, served as the cornerstone for understanding past fluctuations and seasonal variations in rice production. Utilizing this trained ARIMA model, forecasts for future rice yields have been generated. These forecasts extend beyond the historical data, offering insights into the expected trajectory of rice production in upcoming periods. The forecasted values have undergone rigorous validation against actual observed data from corresponding timeframes. This meticulous evaluation process, employing metrics such as mean squared error (MSE), root mean squared error (RMSE), or mean absolute error (MAE), has provided an understanding of the model's accuracy and reliability in predicting rice yields.

3.3 Comparison analysis

This section compares the recommended approach with conventional methods, including *K* closest neighbour (k-NN), support vector machinery (SVM), random forests with Ada boost (RF-AB) random forests with XG-Boost (RF-XGB) enhancement, random forests (RF), gradient boost (GB), and random forests with gradient GB (RF-GB).

The following equation is used to assess assessment index model variables, such as the MAE and RMSE:

$$MAE = \frac{\sum_{i_z=1}^{m_z} |y_{zi_z,f_z} - \overline{y_{zf_z}}|}{m_z} \quad (12)$$

Where y_{zi_z,f_z} is the prediction, $\overline{y_{zf_z}}$ is the true value, and m_z is the total number of data points.

$$\text{RMSE} = \sqrt{\frac{\sum_{i_z=1}^{m_z}(x_{zi_z,f_z} - \overline{x_{zf_z}})^2}{m_z}} \quad (13)$$

Where x_{zi_z,f_z} is the actual observation time series, m_z is the number of non-missing data points, $\overline{x_{zf_z}}$ is the estimated time series, i_z is the variable m_z.

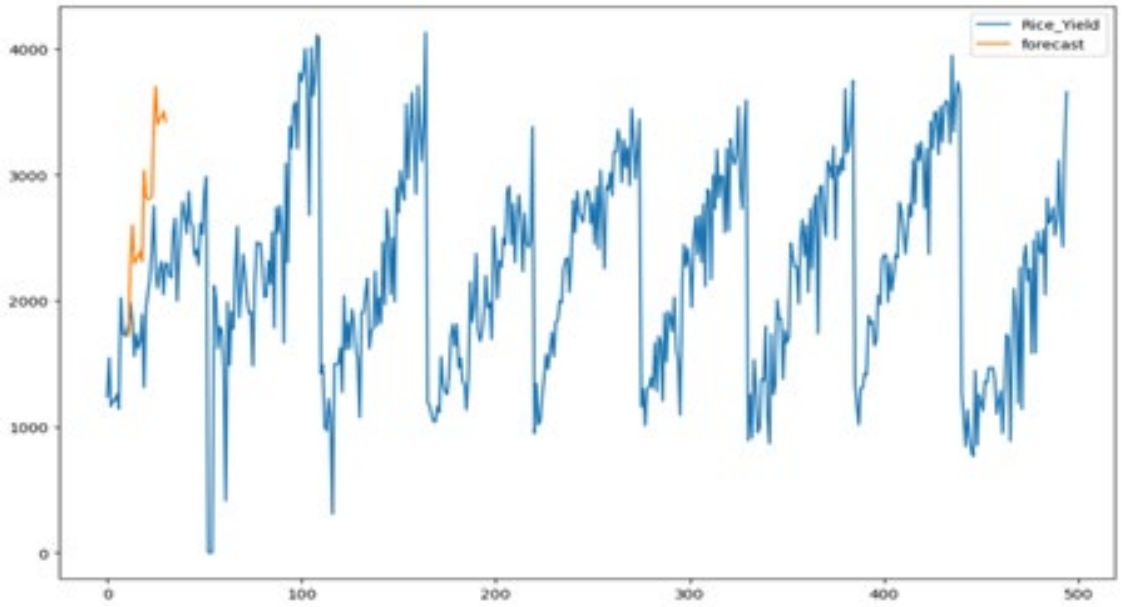

Figure 5 Rice yield vs forecast

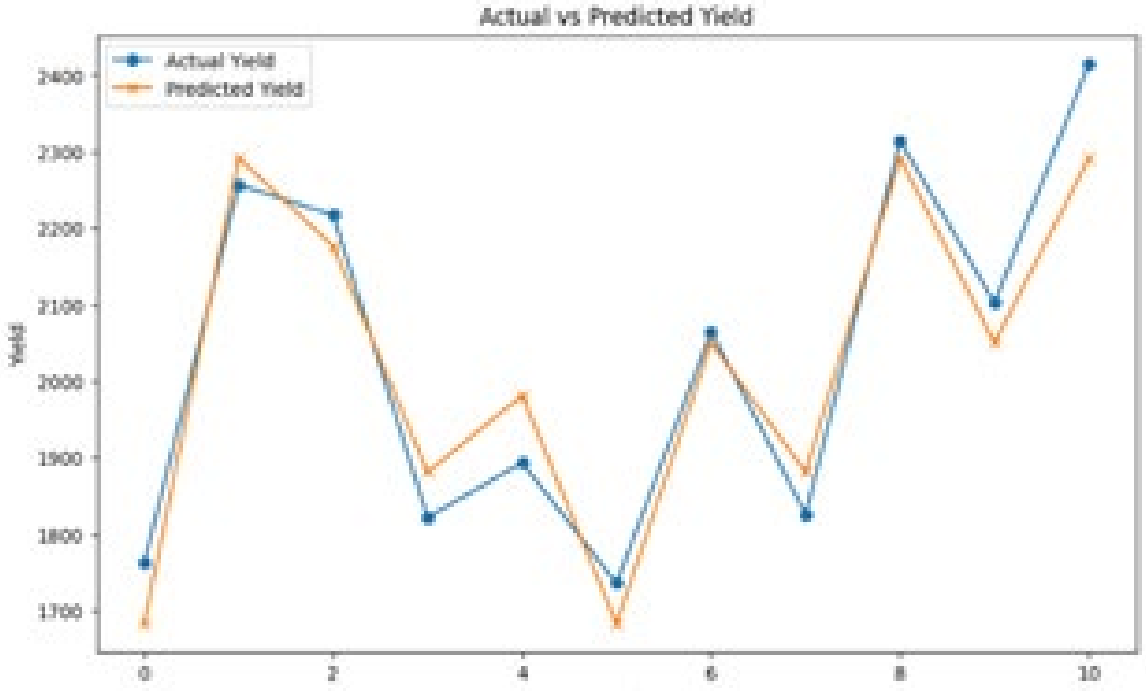

Figure 6 Actual vs predicted yield

3.3.1 Mean absolute error comparison

Figure 7 illustrates MAE of the offered approach. The offered system decreases the fault by incorporating ARIMA model with XGB classifier. Our suggested method in comparison to the baseline RF-AB, RF, GB, RF-XGB, RF-GB, K-NN, and SVM such as 225.9, 232.8, 322.5, 769.9, 330.3, 217.7, 232.9, and 240.0. The MAE of the proposed approach is 44.20. As such, our suggested method works more effectively than the existing methods.

3.3.2 Mean squared error comparison

Figure 8 illustrates MSE of the offered approach. The offered system decreases the error by incorporating ARIMA model with XG Boost classifier. Our suggested method in comparison to the baseline RF-AB, RF, GB, RF-XGB, RF-GB, K-NN,

and SVM such as 93109.5, 187916.1, 1049533.9, 187103.2, 71469.8, 81143.2, and 89389.8. The MSE of the proposed approach is 2591.89. As such, our suggested method works more effectively than the existing methods.

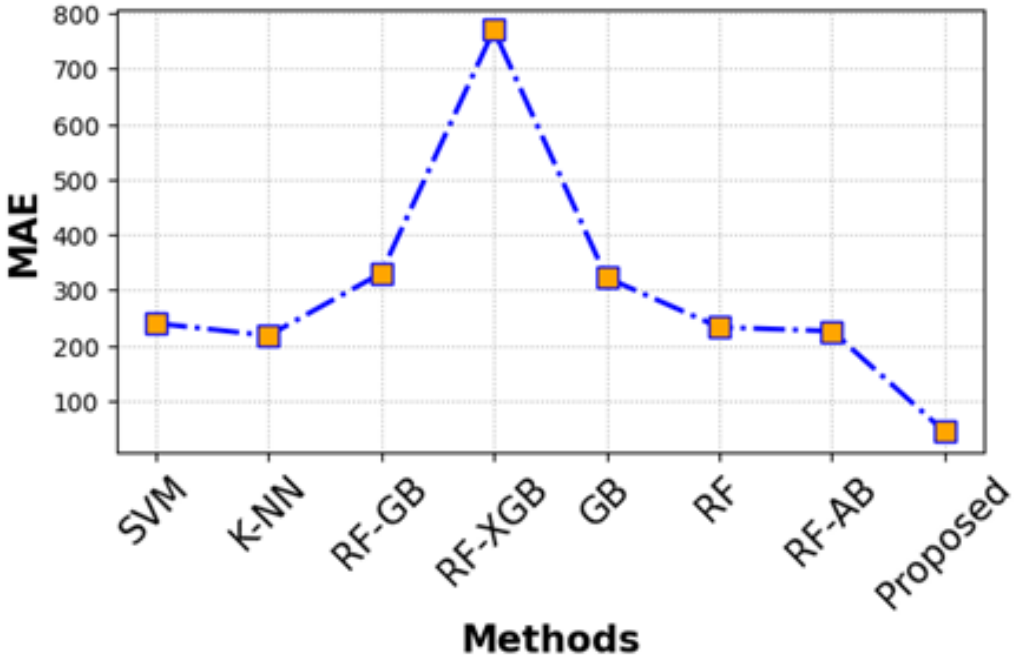

Figure 7 Comparison analysis on MAE

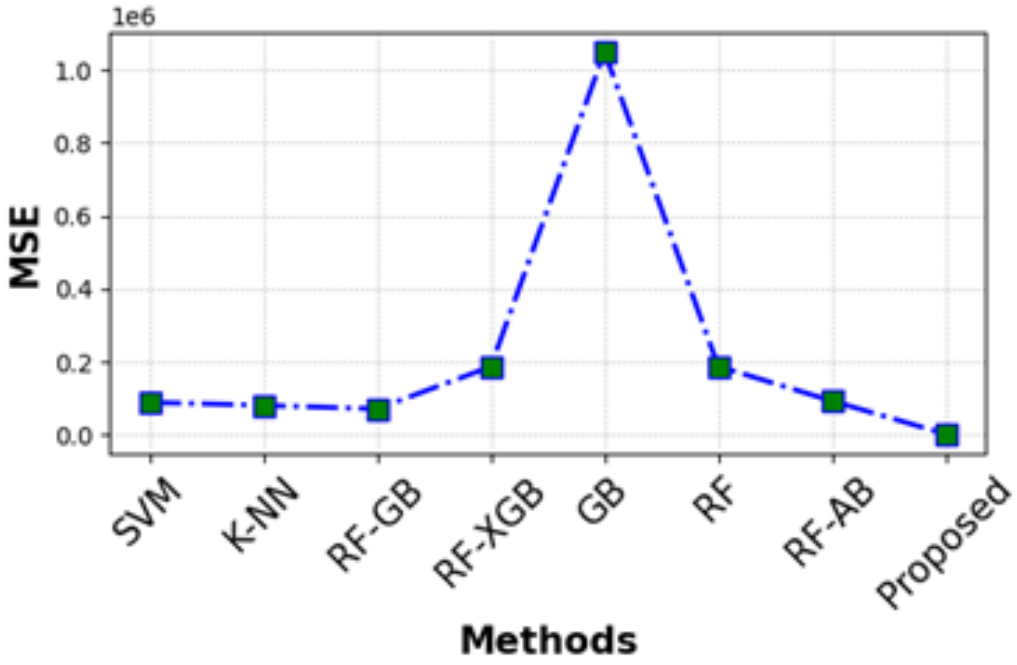

Figure 8 Comparison analysis on MSE

3.3.3 RMSE error comparison

Figure 9 illustrates RMSE of the offered approach. The offered performance decreases the fault by uniting ARIMA model with XG Boost classifier. Our suggested method in comparison to the baseline RF-AB, RF, GB, RF-XGB, RF-GB, K-NN, and SVM such as 282.1, 305.1, 433.0, 812.2, 432.5, 273.4, and 298.9. The RE of the proposed approach is 50.91. As such, our suggested method works more effectively than the existing methods.

3.3.4 R-squared comparison

Figure 10 illustrates R squared Error of the offered approach. The offered system reduces the error by integrating ARIMA model with XG Boost

classifier. Our suggested method in comparison to the baseline RF-AB, RF, GB, RF-XGB, RF-GB, K-NN, and SVM such as 0.853, 0.828, 0.653, 0.2170, 0.868, 0.850, and 0.835. The R squared error of the proposed approach is 0.94. As such, our suggested method works more effectively than the existing methods.

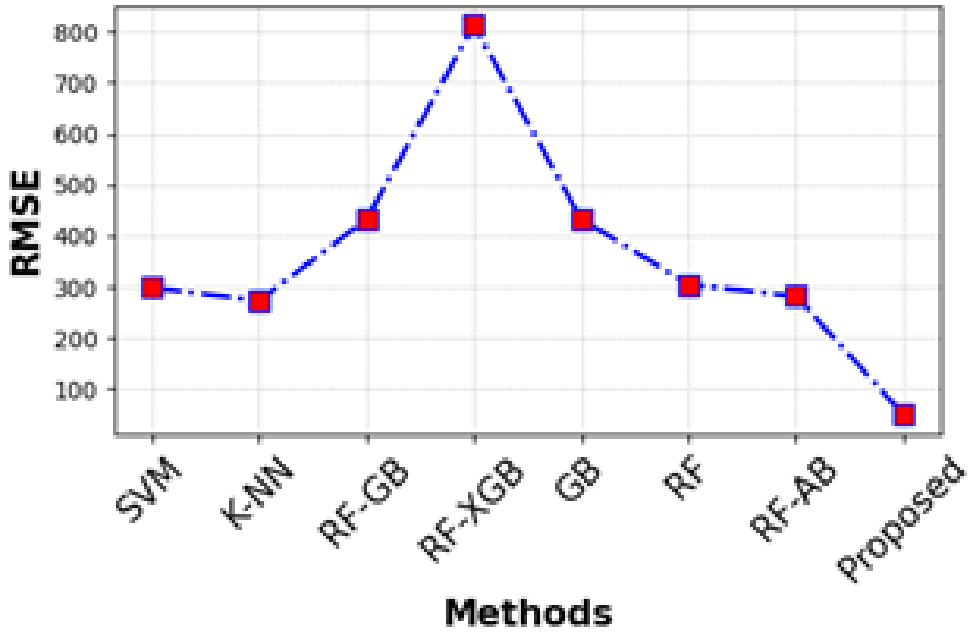

Figure 9 Comparison analysis on RMSE

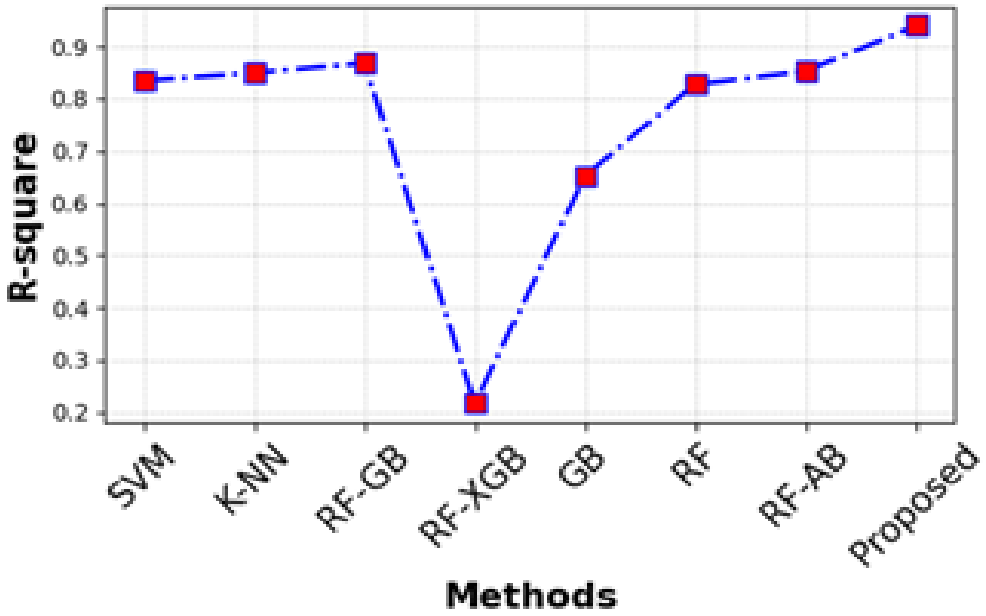

Figure 10 Comparison analysis on R-squared

4 Conclusion

India relies heavily on agriculture for growth and income, particularly rice production. Accurate rice crop yield forecasting is crucial for policy decisions and food supply. Human experience practices often result in low yields. Precision agriculture and smart farming, powered by machine learning and AI, offer automated solutions. However, ensemble techniques face challenges like increased complexity, overfitting, and generalization issues. This work proposes an innovative prediction method to overcome these limitations. Therefore, the proposed method combines hierarchical time series forecasting with ensemble techniques for crop yield prediction, improving accuracy. The dataset is pre-processed, with missing values imputed using k-NN imputation and normalization. The crop yield data is organized into a hierarchical structure, using ARIMA as a model. The

ARIMA typical is trained on training data and validated on test data. The model generates forecasts at the base level, and a top-down stacking ensemble is used to combine ARIMA forecasts with test set data. This approach leverages the strengths of both approaches, resulting in more accurate and robust crop yield predictions.

References

Ajithkumar, B., V. Harithalekshmi, and A. Vysakh. 2021. Rice yield forecasting using principal component regression and composite weather variables. *Journal of Pharmacognosy and Phytochemistry*, 10(2): 595-600.

Apat, S. K., J. Mishra, K. Srujan Raju, and N. Padhy. 2022. State of the art of ensemble learning approach for crop prediction. In *the Conf. on Next Generation of Internet of Things (ICNGIoT 2022)*, 675-685. Odisha, India, 3–4 February.

Attri, S. D., and L. S. Rathore. 2003. Pre-harvest estimation of wheat yield for NW India using climate and weather forecast. *Mausam*, 54(3): 729-738.

Bhadouria, R., R. Singh, V. K. Singh, A. Borthakur, A. Ahamad, G. Kumar, and P. Singh. 2019. Agriculture in the era of climate change: Consequences and effects. In *Climate Change and Agricultural ecosystems: Current Challenges and Adaptation*, eds. K. K. Choudhary, A. Kumar, and A. K. Singh, ch. 1, 1-23. England: Woodhead Publishing.

Cai, Y., K. Guan, D. Lobell, A. B. Potgieter, S. Wang, J. Peng, T. Xu, S. Asseng, Y. Zhang, and B. Peng. 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*, 274: 144-159.

Chandraprabha, M., and R. K. Dhanraj. 2023. Ensemble Deep Learning Algorithm for Forecasting of Rice Crop Yield based on Soil Nutrition Levels. *EAI Endorsed Transactions on Scalable Information Systems*, 10(3): 2610.

Das, B., B. Nair, V. K. Reddy, and P. Venkatesh. 2018. Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. *International Journal of Biometeorology*, 62(10): 1809-1822.

Deryng, D., D. Conway, N. Ramankutty, J. Price, and R. Warren. 2014. Global crop yield response to extreme heat stress under multiple climate change futures. *Environmental Research Letters*, 9(3): 034011.

Do Land, H. M 2009. The resource outlook to 2050. Available at: https://projects.mcrit.com/foresightlibrary/attachments/resource_outlook_2050.pdf.

FAOSTAT. 2022 FAOSTAT Online Database. Available online: http://faostat3.fao.org/browse/Q/QC/E..

Feng, L., Y. Wang, Z. Zhang, and Q. Du. 2021. Geographically and temporally weighted neural network for winter wheat yield prediction. *Remote Sensing of Environment*, 262: 112514.

Gupta, R., and A. Mishra. 2019. Climate change induced impact and uncertainty of rice yield of agro-ecological zones of India. *Agricultural Systems*, 173: 1-11.

Iniyan, S., and R. Jebakumar. 2022. Mutual information feature selection (MIFS) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression (MSER). *Wireless Personal Communications*, 126(3): 1935-1964.

Kamath, P., P. Patil, S. Shrilatha, and S. Sowmya. 2021. Crop yield forecasting using data mining. *Global Transitions Proceedings*, 2(2): 402-407.

Kundu, S. G., A. Ghosh, A. Kundu, and G. Gp. 2022. A ML-AI enabled ensemble model for predicting agricultural yield. *Cogent Food & Agriculture*, 8(1): 2085717.

Li, L., B. Wang, P. Feng, H. Wang, Q. He, Y. Wang, D. Liu, Y. Li, J. He, H. Feng, G. Yang, and Q. Yu. 2021. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. *Agricultural and Forest Meteorology*, 308: 108558.

Ma, Y., Z. Zhang, Y. Kang, and M. Özdoğan. 2021. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment*, 259: 112408.

Nain, G., N. Bhardwaj, P. M. Jaslam, and C. S. Dagar. 2021. Rice yield forecasting using agro-meteorological variables: A multivariate approach. *Journal of Agrometeorology*, 23(1): 100-105.

Oikonomidis, A., C. Catal, and A. Kassahun. 2022. Hybrid deep learning-based models for crop yield prediction. *Applied Artificial Intelligence*, 36(1): 2031822.

Palanivel, K., and C. Surianarayanan. 2019. An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology*, 10(3): 110-118.

Paudel, D., H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis. 2021. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187: 103016.

Rajak, R. K., A. Pawar, M. Pendke, P. Shinde, S. Rathod, and A. Devare. 2017. Crop recommendation system to maximize crop yield using machine learning technique.

International Research Journal of Engineering and Technology, 4(12): 950-953.

Ray, D. K., N. D. Mueller, P. C. West, and J. A. Foley. 2013. Yield trends are insufficient to double global crop production by 2050. *PloS One*, 8(6): e66428.

Shahhosseini, M., G. Hu, I. Huber, and S. V. Archontoulis. 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports*, 11(1): 1606.

Singh, A. K., A. Vashisth, V. K. Sehgal, A. Goyal, H. Pathak, and S. S. Parihar. 2014. Development of multi stage district level wheat yield forecast models. *Journal of Agricultural Physics*, 14(2): 189-193.

Singha, M., J. Dong, G. Zhang, and X. Xiao. 2019. High resolution paddy rice maps in cloud-prone Bangladesh and Northeast India using Sentinel-1 data. *Scientific Data*, 6(1): 26.

Soora, N. K., P. K. Aggarwal, R. Saxena, S. Rani, S. Jain, and N. Chauhan. 2013. An assessment of regional vulnerability of rice to climate change in India. *Climatic Change*, 118: 683-699.

Tilman, D., C. Balzer, J. Hill, and B. L. Befort. 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, 108(50): 20260-20264.

Yesugade, K. D., A. Kharde, K. Mirashi, K. Muley, and H. Chudasama. 2018. Machine learning approach for crop selection based on agro-climatic conditions. *Machine Learning*, 7(10): 103-106.

Zabel, F., C. Müller, J. Elliott, S. Minoli, J. Jägermeyr, J. M. Schneider, J. A. Franke, E. Moyer, M. Dury, L. Francois, C. Floberth, W. Liu, T. A. M. Pugh, S. Olin, S. S. Rabin, W. Mauser, T. Hank, A. C. Ruane, and S. Asseng. 2021. Large potential for crop production adaptation depends on available future varieties. *Global Change Biology*, 27(16): 3870-3882.

Zhang, G., X. Xiao, C. M. Biradar, J. Dong, Y. Qin, M. A. Menarguez, Y. Zhou, Y. Zhang, C. Jin, J. Wang, R. B. Doughty, M. Ding, and B. Moore III. 2017. Spatiotemporal patterns of paddy rice croplands in China and India from 2000 to 2015. *Science of the Total Environment*, 579: 82-92.

Zhang, G., X. Xiao, J. Dong, F. Xin, Y. Zhang, Y. Qin, R. B. Doughty, and B. Moore III. 2020. Fingerprint of rice paddies in spatial–temporal dynamics of atmospheric methane concentration in monsoon Asia. *Nature Communications*, 11(1): 554.