International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2) 2026 pp. 48-55

https://doi.org/10.70917/ijcisim-2026-2150
Article

Enhancing Wheat – Yield Forecasting with Ensemble Random Forest Machine Learning Technique

D.K.Adlin Femi[1], D.Sheeba Singh[2]
[1]Department of Mathematics, Noorul Islam Centre for Higher Education, Kumaracoil, Thuckalay, kanyakumari, India.
Register No:1122916101, Email: adlinfemi9674@outlook.com
[2]Department of Mathematics, Noorul Islam Centre for Higher Education, Kumaracoil,Thuckalay, kanyakumari, India.

Abstract: Economy and food security of our country depends heavily on wheat, therefore precise yield forecasts is crucial for resource management and planning. Although traditional methods, such remote sensing and manual field surveys have been widely employed, it is still unclear how well they capture yield variance throughout various growth phases. Predicting crop yields is essential for increasing productivity, controlling risks, guaranteeing food security, and boosting agriculture's general sustainability. The Random Forest (RF) machine learning approach was assessed in this study for its capacity to forecast wheat production of the country in the period 2023-2030, responses to soil and climate-related parameters with Multiple Linear Regression (MLR). Evaluation measures included Error in Root Mean Square (RMSE) also Absolute Mean Error (AME), with RF. This finding indicates, RF is better than Existing Markov Chain method at predicting wheat production.
Keywords: Machine intelligence, sustainability, and forecasting wheat production

1. INTRODUCTION

One of the most extensively farmed cereals in the twenty-first century, wheat provides a disproportionate amount of protein and calories to the global food supply. The precision of crop output forecasts is crucial in the grain circulation market in order to prevent famine and provide food security. Fertilization efforts also aided by forecasting the amount of food a crop will produce. Accurate crop output forecasts are useful for business, national food policy, and security planning. One of the most extensively farmed cereals in the twenty-first century, Wheat contributes out of proportion quantity of calories and protein to the world's food source. The precision of crop output forecasts is crucial in the grain circulation market in order to prevent famine and provide food security. Fertilization efforts can also be aided by forecasting the amount of food a crop will produce. Accurate crop output forecasts are useful for business, national food policy, and security planning. A steady and adequate national food supply is facilitated by the government's ability to make knowledgeable choices on exports and imports thanks to accurate crop estimates. These estimates are especially crucial for nations like Pakistan, which are dealing with serious issues brought on by population increase and the effects of climate change. These projections are the main source of trade surpluses and GDP growth, guaranteeing the availability of food in both rural and urban areas. Many planners and administrators recognize the critical relevance of timely and precise farm information. Because Corn, wheat, rice, sugarcane, cotton, and other basic crops are so important, Legislators always look for greatest accurate also current data regarding production.

2. LITERATURE REVIEW

Wheat is among the world's most significant staple crops, as well as ongoing inherent development is necessary to fulfill the requirements of the worlds continuously expanding demographic. Conventional breeding produced the high-yielding wheat cultivars that exist today; new advances in genetic engineering should improve traditional procreation to increase produce even so further. Advances in molecular breeding and genome sequencing have accelerated the rate of gene discovery, requiring the creation of incredibly dependable and efficient modification mechanisms. [1] The focused alteration of the genome will need the economical conveyance of nucleases that are Sequence-dependent like zinc fingers (ZFNs), transcription-activating –such as effect nuclease (TALENs), also RNA- directed designer nuclease like Cas9-CRISPR. However, like with other cereal crops, genetic influences have also impeded wheat transformations.

Just to meet every rising demand, wheat production must rise significantly (& gt; 40% by 2020). Due to the small gene pool available, traditional plant breeding techniques are unlikely to produce this growth. In addition

to being desired, the use of recombinant technology to increase wheat production and quality may create new opportunities. [2] Even though gene-transformation techniques for enhancing these characteristics have advanced significantly, this remains a significant barrier to plant biotechnology. The introduction of little to no inter generic DNA, the necessity to develop elite wheat varieties without selectable markers, and the commercial and social concerns around genetically engineered food products are some of the challenges in realizing the full promise of the genomic age in wheat breeding.

A fast, accurate, and trustworthy large-scale crop yield estimate is higher important more important for crops administration, evaluating sustenance safety, trading sustenance, and formulating strategy in order to address every obstacle posed by weather variability, growing populations, also sustenance requisition. In this study, we compared by combining three deep learning (DL) models—DNN (deep neural networks), 1D-CNN (1D Convolutional neural networks), and LSTM (long short-term memory networks)—with a conventional machine learning approach (random forest, RF) [3] to forecast crop yields, the GEE platform integrates publicly accessible data, such as climate, satellite, and soil characteristics. The findings demonstrated that changes in winter wheat output over all county-years could be captured by all four models.

Given that Beccari found gluten in the eighteenth hundred year, wheat producers, millers, and bakers have looked for comparatively easy ways to anticipate wheat quality, especially during the breeding process. Since endosperm proteins are the primary determinant of wheat quality, the prediction is typically based on these proteins. [4] Early studies focused on relationships between macro fractions, such as the ratio of acetic acid soluble proteins to gliadin and glutenin. Subsequently, research focused on gliadin polypeptides, the genes that code for these polypeptides, and variety.

Using spectral and imaging data, systems that forecast the produce of Milling Yield (MY %) were developed obtained from 996 wheat samples. The information included spectrum of the close to infrared reflections the relevant large cereal and multispectral photos of individual wheat grains. [5] Sets of features that were taken data and roughly divided into spectral and picture characteristics (sample grain mean and standard deviations attributes), such as the color, shape, and morphology of the grains, were used to describe the samples. Deep learning was used in image feature extraction to classify grain orientation, estimate the proportionate sizes in the embryo and endosperm, also estimate the depth of grain creases. Regression models across qualities were chosen using extracted features as independent variables identification. High-molecular-weight glutenin subunits constitute the basis for more recent predictions. Despite the fact that several nations have effectively used the latter prediction in wheat breeding, there is little link between expected and actual quality.

The random forest algorithm was first presented by L. Breiman, also its proven being highly successful broadly applicable classification also regression method. The approach, which blends several arbitrarily trees of choice also mean their forecasts, has been shown remarkable execution when the quantity of factors are much greater compared to the quantity of findings. [6] It’s also sufficiently diversified to be used in big-scale issues, returns metrics of varying relevance, and is readily able to adjust to different ad hoc gaining knowledge difficulties. The most current theoretical and methodological advancements for random forests are reviewed in this article. The algorithm mathematical motivations are emphasized, with particular focus on parameter selection, the re-sampling technique, and variable importance measurements.

The values of a randomly selected vector with the identical dissemination also Free-standing sampling tree within a random forest ascertain the ideals of all trees. This is accomplished by merging tree predictors. [7] A forest's generalization error approaches a limit as its tree count rises. A forest of tree classifiers' generalization error is determined by each tree's strength and the correlation between them. More noise-resistant rate of error and comparable to Adaboos Y. Freund & R. Schapire are obtained by splitting each node using a random selection of features. Internal estimates are used to track error and display the reaction of raising the number.

A collective classifier referred to as a random forest (RF) classifier generate several selection trees utilizing a randomly selected subset of instructing variables and samples. In the realm of remote sensing, this classifier has grown in prominence owing to the precision of its categories. The primary purpose of this task aimed to examine the application of RF classifiers within detecting remotely. [8] According to this preview, the RF classification is fast, insensitive to over fitting, and capable of handling large data dimensionality and multi colinearity. However, the sample design has an impact on it. There are numerous applications for the RF classifier variable important (VI) measurement, such as identifying greatest pertinent multisource distant sensing and geographical information by reducing the number of dimension.

A design for machine learning called a random forest is utilized for categorization and forecasting. For machine learning method and artificial intelligence designs to be trained effectively, a significant quantity of excellent information is required. System performance data is crucial for improving algorithms, increasing hardware and software efficiency, evaluating user behaviour, and supporting decision-making, problem-solving, predictive modelling, and pattern detection, all of which lead to increased efficacy and accuracy. [9]

Using a variety of data collection and processing techniques improves problem-solving accuracy and creativity. The use of data analysis results for problem-solving, decision-making, predictive modelling, and pattern recognition is made possible by the use of varied approaches in interdisciplinary research, which also simplifies the research process and encourages creativity.

The Random Forest algorithm is a innovative machine intelligence also pairing technique. To generate Random Forest, several tree structure classifiers are combined. [10] There are a lot of positive Random Forest Personages. Random Forest been widely utilized in regression, categorization and forecasting. Random Forest has numerous advantages over standard algorithms. As a result, Random Forest has a wide range of applications.

The overuse of groundwater for wheat irrigation has been made worse by changes in precipitation patterns brought on by climate change. Few researches have explicitly evaluated the consequences of sowing wheat on shelf of groundwater (GWS), despite this fact that numerous studies have looked at the effects in wheat agriculture using GWS. [11] To evaluate the impact of region wheat on GWS throughout China, proposing a method for wheat discovery of subsiding effects utilizing time-based distant sensing photos. The GWS difference within the self-adjusting window that moves between wheat region also surrounding non-wheat area is used to calculate the subsiding magnitude of the WSED (Depending at the properties of the information at the core pixel point, sliding window's size and position can be automatically changed.

In order to address issues brought on by farmers' advanced age and lack of heirs, information and communication technology (ICT) is necessary in the agricultural sector. In example, current agricultural support systems make extensive use of environmental sensors and cameras to facilitate data collection. [12] Utilizing the information to machine learning and innovations for data mining currently anticipates the spread of smart farming, even though the customary role among these setup, unsophisticated environmental keeping an eye on and control. In order to accurately forecast complex water strain out of two types of information — environmental and plant image data—we thus offer a unique multiple models SW-SVR method has support vector regression technique based on sliding windows. The suggested approach uses a SW-SVR also a deep neural network (DNN) as multi-model attribute retrieval.

A unique time-variant weighting technique that integrates energetic period distortion separations and model of sliding windows is developed along with predicting to address this situation out of temporal delay in single projected outcomes. The observed and expected time series micro-level statistical features can be model by using a sliding window. However, the DTW algorithm introduction allows for the measurement of both the similarity and the separation within what was seen and predicted time series. [13] In combination forecasting, the combination of a movable window and DTW distant may offer an innovative apparatus that fully captures overall incomplete quantitative properties of temporal intervals. The created predicting combinations model viability and validity are demonstrated by two numerical studies.

Very damaging disease that negatively affects the whole expansion of wheat plant is wheat scab. In order to stop wheat scab from spreading, it is essential to quickly assess the field wheat scab levels. However, manual observation is time-consuming and ineffective. According to recent studies computer view oriented techniques improve productivity within the area. [14] This analysis suggested rotation detector and Swin classifier approach for wheat scab level prediction. The analysis included revolving deterrent for wheat (RWD) connection for wheat head detection in order to reduce background interference. To increase accuracy, the RWD system used the Intersection of Kalman filters over Union (KFIoU) to estimate the inclination. To distinguish between healthy and sick wheat heads, the Swin Wheat Classifier (SWC) system were utilized.
Accurately predicting the way wheat costs are going to fluctuate can become a useful tool. Most of authored work do not always do out-of-sample testing and instead apply customary estimating algorithms to wheat cost temporal sequence. [15] Using just historical temporal sequel data, this research examines five modelling techniques to wheat cost forecasts. By taking into account a moving and expanding time frame that will outline content required to set the designs characteristics and the data used to forecasts made outside the specimen, the model performance is evaluated solely using out- of-sample data.

3. Proposed Method

Figure1 depicts the complete study workflow. It provides a detailed description of the data gathering phase of this investigation. Crop masks were used to filter all spatial data so as to determine wheat regions in the subsequent stage. Along with the preprocessed data, training information is given to predictive model, and the wheat yield is predicted and comparison is made with the statistical yield data.

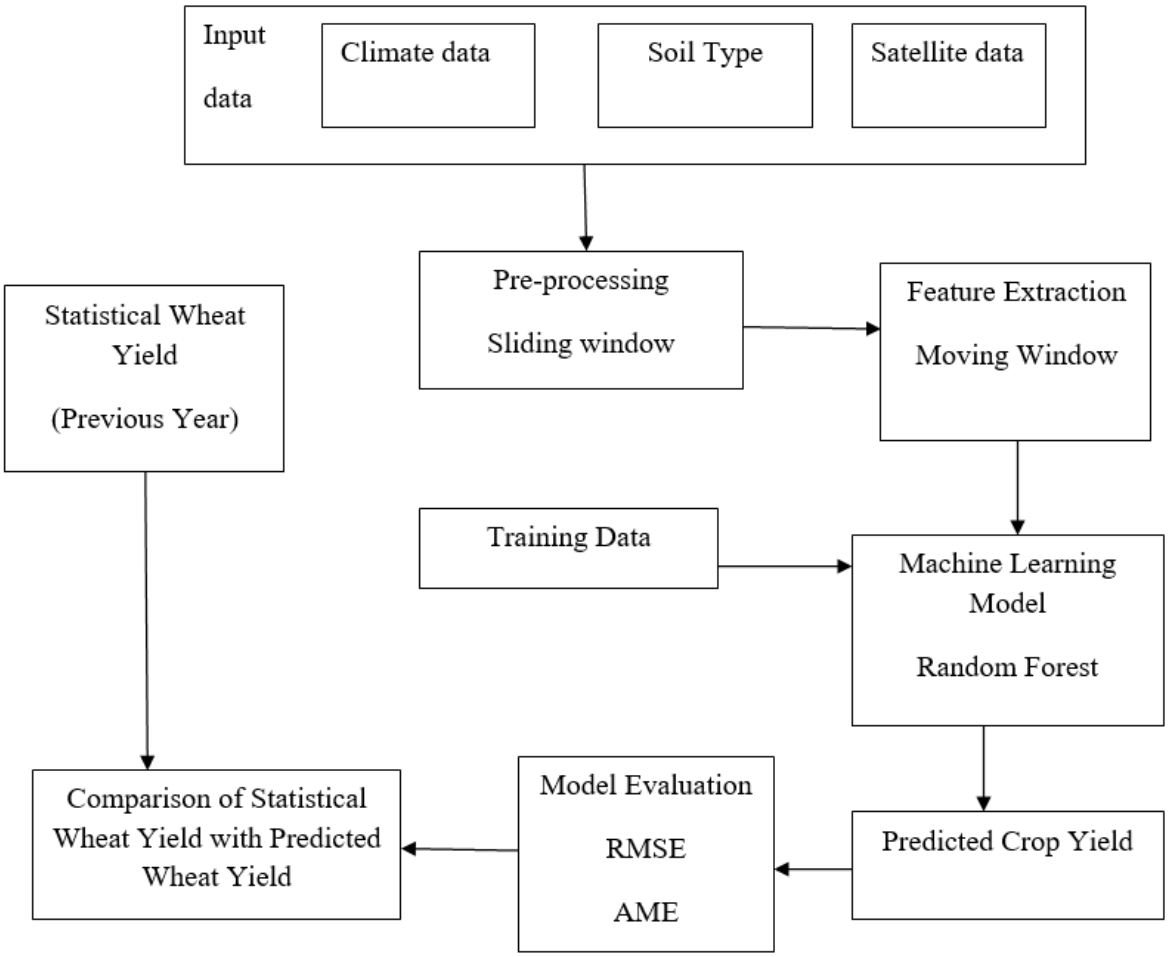

Fig.1 Frame work of the Proposed RF Model

3.1 Input Dataset

The Soil data utilized in this analysis for model testing and training came from a variety of sources and covered the years 1980 to 202025. Even with consistent environmental and management circumstances, crop growth and development are impacted by the differences in soil properties between fields, which eventually influence the final yield. The National Meteorological Agency provided the climate data. The lowest possible thermostat, typical thermostat, highest possible thermostat, lowest possible shortfall in vacuum pressure, and highest possible vacuum pressure short fall are among meteorological data that were taken from this dataset. The satellite data are accessed through the Google Earth Engine Platform

3.2 Pre- Processing

A crucial stage in getting datasets ready for machine learning models is data preparation. The execution of machine learning models is significantly impacted by missing data. To solve this problem, a variety of strategies are used, including the methods fill in the gaps in datasets to preserve the model's integrity. For example, one-hot encoding has demonstrated accuracy and robustness in situations with high missing rates.

Sliding Window Technique

Sliding window technique finds anomalies by recognizing values that fall outside of a normal pattern distribution. Sliding window perform the following function.

- Determine the total of initial L items not in m terms utilizing a linear sequence, storing outcome in the total changing window.
- Next, keep track of maximum sum while moving across the array in a linear fashion until it comes to final.
- Simply subtract to find present total enclose of L items, Add the final element of the present block to the first element from the preceding block.

3.3 Feature Extraction Using Moving window Technique

Moving Window feature extraction is a method in machine learning where a fixed size window slides

over a continuous data stream, splitting it into overlapping segments. As the window moves, it captures temporal patterns and trends, converting raw data into a more informative representation. Common extracted features include statistical metrics like mean and variance, as well as spectral features from Fourier transforms. This method is widely used in agricultural monitoring to improve model performance by highlighting relevant patterns.

$$Mean\ (\mu) = \left(\frac{1}{N}\right) * \sum(x\_i) \quad (1)$$

$$Variance\ (\sigma^2) = \frac{1}{N} * \sum(x_i - \mu)^2 \quad (2)$$

$$\text{FFT}\ (X\_K) = \sum(x_n * e^{-\frac{i2\pi kn}{N}}) \quad (3)$$

Where, x_i – Data Points in the Window

N - Window size

x_n – Data Points

k – Bin Index

3.4 Training Dataset

A machine learning or computer vision algorithm or model is taught to process information using training data, which is the first training dataset. Algorithmic models learn from and comprehend the information by using the raw data. As they come across new data and expand on what they have learnt from the earlier data, these models continue to improve their performance increasing their confidence and decision-making —. Because the development, performance, and accuracy of any model are significantly impacted by the quality of the training data, high-quality training data is the cornerstone of successful machine learning. Because the volume the caliber of the label training information directly alter how well model learns to identify the result it was intended to detect, training data is just as important to the success of a production-ready model as the algorithms themselves.

3.5 Random Forest Predictive Model

Wheat yield was predicted utilizing a Random Forest (RF) regression model. RF has a collective machine learning technique which improves prediction accuracy and demonstration by integrating many decision-making trees. Every tree is trained using an arbitrary segment of data also predictions are generated by averaging the outcomes of all the trees. This method is resistant to over fitting. To determine how much the model depends on a specific feature while making predictions, feature importance and partial dependence plots were developed. When all other factors are held constant, the partial dependence plots show how the model's predictions vary with regard to the most significant feature.

Every tree in the RF, regression is constructed using an arbitrary choice of variables and samples from the dataset. Each decision tree then creates a Multiple Linear regression model and generates a prediction, each of which represents a sub-regression model. Finally, all of the forecasts from every decision making tree are averaged to provide the ultimate forecast.

Matrices Evaluation of Multiple Linear Regression

Absolute Mean Error (AME) and Error in Root Mean Square (RMSE) were utilized to assess the system performance in making predictions on the test set. These are typical evaluation measures for crop prediction; higher model performance is indicated by lower values.

The metrics were computed in the following manner

$$\boldsymbol{AME = \frac{1}{n}\sum_{j=1}^{m} |x_i - \hat{x_i}|} \quad (4)$$

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{m}(x_i - \hat{x_i})^2}} \quad (5)$$

Where, $\boldsymbol{m}$ *- quantity of observations*

$\boldsymbol{x_i}$*- The target variable's observed value*

$\boldsymbol{\hat{x_i}}$ *- The target variable's anticipated value*

4. RESULT AND DISCUSSION

The study shows that the Random Forest (RF) model is a useful instrument for wheat yield prediction. Using test data, the RF model predicted wheat yield and calculate Absolute Mean Error (AME) and Error in Root Mean Square (RMSE). The same dataset is utilized to instruct and assess the RF system. Two calculated metrics help gauge the system accuracy also highlight region of upgrades. The RF model's predictions fit the data quite well, as shown in Table 1 comparing the statistical and predicted values.

RF Model	RMSE	AME	Predicted Wheat Yield	Statistical Wheat Yield
Climate	164.70	126.89	13.65	10.78
Soil Type	145.56	146.89	15.67	12.65
Satellite Data	135.78	167.35	17.67	14.89

Table 1 Model Evaluation Parameters

The bar chart shown in Fig. 2 compares the predicted Wheat yield from a proposed Random Forest (RF) model with actual statistical wheat yield. The graph indicates the predicted yield is higher than the statistical yield.

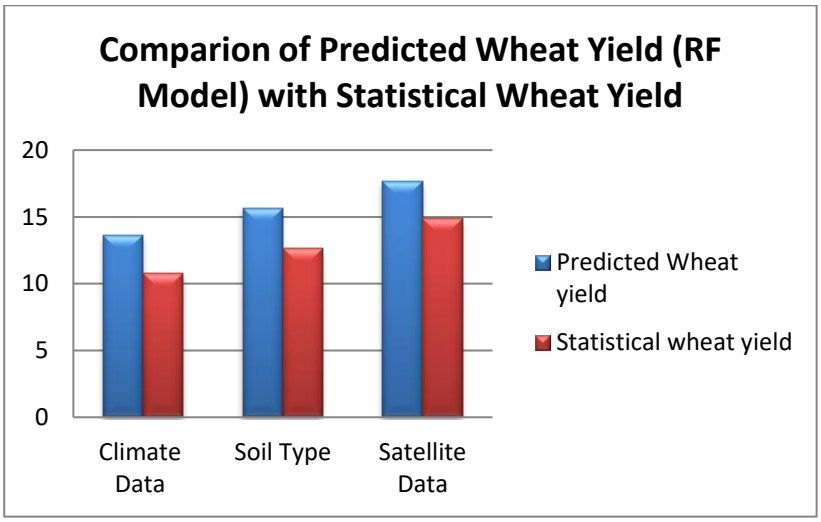

Fig.2 Comparison of RF Model's prediction with statistical Prediction

Years	Predicted Wheat Yield Using RF Model	Existing Markov Chain for wheat yield Prediction
2023	876.89	782.84
2024	886.79	782.89
2025	889.78	782.94
2026	897.78	779.83
2027	896.90	773.96
2028	976.76	769.07
2029	843.78	763.02
2030	945.27	759.14

Table 2 Wheat Yield Forecast

The Table 2 represents the Wheat – yield forecasts for the years 2023 to 2030. The Proposed RF model is compared with Existing Markov Chain Model.

The chart shown in Fig. 3 compares the predicted wheat yield for the period 2023-2030. This visual indicates that the RF algorithm provides systematically larger yield estimates, suggesting it predicts better performance

or higher production potential compared with the traditional Markov chain technique for Wheat Yield Forecasting

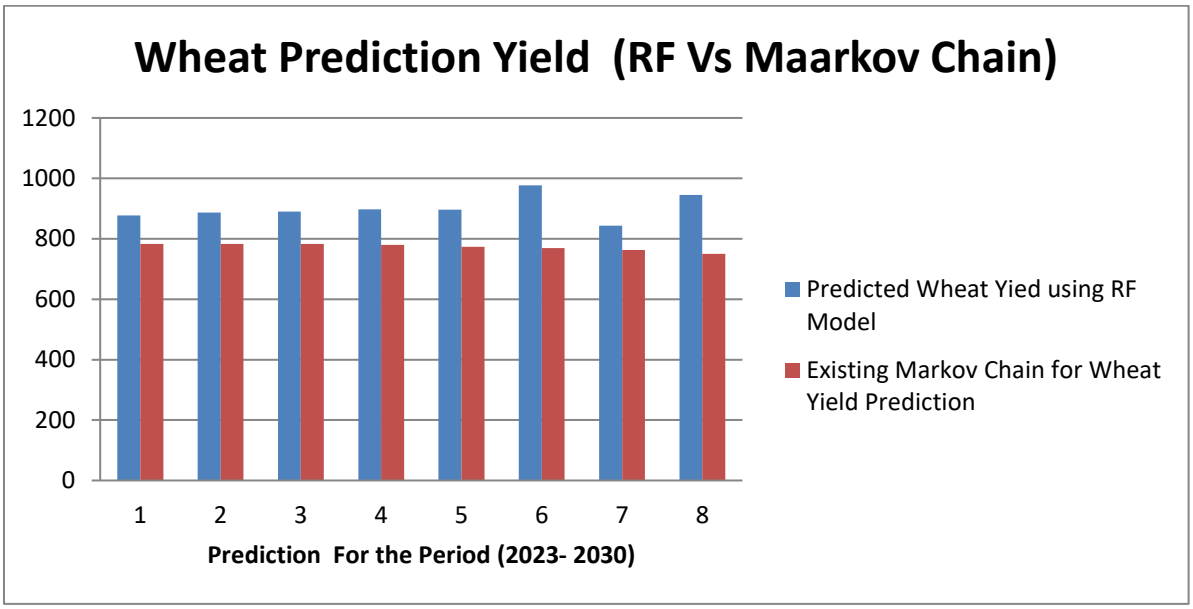

Fig. 3 Comparison of RF VS Markov Chain Model

5. CONCLUSION

The study concludes the precise wheat- yield forecasting is vital for resource management and food security in the country, because wheat heavily influences the economy. Traditional method like field survey is used but do not adequately capture yield variations across growth phases. The research evaluates the Random Forest (RF) machine learning approach for predicting wheat prediction in the 2023-2030 period, using soil, climate and satellite related parameters combined with Multiple Linear Regression (MLR). Execution from the suggested system is established using Error in Root Mean Square (RMSE) and Absolute Mean Error (AME). The outcomes demonstrates, RF model outperforms existing Markov Chain method in forecasting wheat yields, indicating its superiority for improving agricultural sustainability and ensuring food security.

References

1. Shrawat, Ashok K., and Charles L. Armstrong. " Development and application of genetic engineering for wheat improvement. " Critical Reviews in Plant Sciences 37, no. 5 (2018): 335-421.
2. Bhalla, Prem L. "Genetic engineering of wheat–current challenges and opportunities." TRENDS in Biotechnology 24, no. 7 (2006): 305-311.
3. Cao, Juan, Zhao Zhang, Yuchuan Luo, Liangliang Zhang, Jing Zhang, Ziyue Li, and FuluTao. " Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine." European Journal of Agronomy 123 (2021): 126204.
4. Lásztity, Radomir. "Prediction of wheat quality-Success and doubts." Periodica polytechnicachemical engineering 46, no. 1-2 (2002): 39-49.
5. Assadzadeh, Sahand, Cassandra K. Walker, Linda S. McDonald, and Joe F. Panozzo."Prediction of milling yield in wheat with the use of spectral, colour, shape, and morphological features." Biosystems Engineering 214 (2022): 28-41.
6. Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25, no. 2 (2016): 197-227.
7. Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.
8. Belgiu, Mariana, and Lucian Drăguţ. "Random forest in remote sensing: A review ofapplications and future directions." ISPRS journal of photogrammetry and remote sensing 114(2016): 24-31.
9. Salman, Hasan Ahmed, Ali Kalakech, and Amani Steiti. "Random forest algorithm overview." Babylonian Journal of Machine Learning 2024 (2024): 69-79.
10. Liu, Yanli, Yourong Wang, and Jian Zhang. "New machine learning algorithm: Random forest." In International conference on information computing and applications, pp. 246-252. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
11. Fan, Lingling, Shi Chen, Lang Xia, Yan Zha, and Peng Yang. "Assessing the effects of wheat planting on groundwater under climate change: a quantitative adaptive sliding window detection strategy." Atmosphere 15, no. 12 (2024): 1501.
12. Kaneda, Yukimasa, Shun Shibata, and Hiroshi Mineno. "Multi-modal sliding window-based support vector regression for predicting plant water stress." Knowledge-Based Systems 134 (2017): 135-148.

13. Tao, Zhifu, Qinghua Xu, Xi Liu, and Jinpei Liu. "An integrated approach implementing sliding window and DTW distance for time series forecasting tasks." Applied Intelligence 53, no. 17 (2023): 20614-20625.
14. Zhang, Dongyan, Zhipeng Chen, Hansen Luo, Gensheng Hu, Xin-Gen Zhou, Chunyan Gu, Liping Li, and Wei Guo. "Predicting wheat scab levels based on rotation detector and Swin classifier." Biosystems Engineering 248 (2024): 15-31.
15. Dias, Joana, and Humberto Rocha. "Forecasting wheat prices based on past behavior: comparison of different modelling approaches." In International conference on computational Science and its applications, pp. 167-182. Cham: Springer International Publishing, 2019.