Intelligent Systems Research and Applications Journal

www.israj.org
eISSN: 3120 - 0826

Data-Driven Predication of Wheat Production across Africa, the Middle East, and Oceania using Machine Learning Techniques

Naman[1], Anju[1], Paras Dahiya[1], Suraj Arya[1],*, Ayodeji Olalekan Salau[2]

[1] Department of Computer Science and Information Technology, Central University of Haryana, Mahendergarh, Haryana, India
[2] Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ekiti State, Nigeria

ARTICLE INFO

Article history:
Received 3 March 2026
Received in revised form 14 May 2026
Accepted 25 June 2026
Available online 3 July 2026

Keywords:
Wheat Production; Machine learning; XGBoost

ABSTRACT

Wheat is a crop that is consumed and produced globally. It is used in many food items in our daily life. Therefore, its widespread use globally and its share in food grains makes its extremely important. Therefore, considering the importance of the wheat crop, it is very important to estimate its production level in advance so that equitable availability of this important grain can be ensured at the global level. This study addresses the limitations of traditional statistical models by incorporating advanced machine learning techniques capable of capturing nonlinear relationships in agricultural data. Experimental results demonstrate that XGBoost outperforms other models with the accuracy level 99.99%. and an RMSE of 18.37 and MAE of 13.96, indicating high predictive accuracy. The findings highlight the potential of machine learning-driven approaches for supporting data-informed agricultural planning and policy formulation in regions with high production variability. This framework will not only help predict wheat production in advance but also serve as a useful tool for agricultural policymaking.

1. Introduction

Agriculture plays a very important role in the economies of many nations globally. It provides food, employment, and opportunities for a majority of the citizens of a country, and supports the growth and development of the economy of these countries. There are a number of factors present, such as climatic variability, soil fertility, technological innovations, and government policies that affect the needs of the agricultural sector of these countries from year to year. For these, it is a very difficult prediction to make concerning the quantity of agricultural products that will be required for food production, since its forecast affects the planning that must be exercised for their consumption and the stability of the economic affairs of these nations. Egypt, Ethiopia, and Algeria are the foremost countries, as regards the production of wheat in Africa, Turkey (21 million tons in 2024), Iran (14 million tons in 2024), and Iraq (3.5 million tons in 2024) in Middle East, Australia (41 million tons in 2023), New Zealand (much less than Australia (production is minor in comparison) in Oceania, and their total production constitutes the major portion of that produced globally. But

* *Corresponding author.*
E-mail address: surajarya@cuh.ac.in
https://doi.org/10.59543/f2yazb53

Africa produces a large part of the wheat consumed in the nation as a whole, but is nevertheless a net importer of the grain in order to keep in line with the consumption of its several nations, which is on account of the growing population density and change of food consumption [1]. The general cry is against the production of more food, and this is being tried through better agricultural methods, better research, and better facilities. The plea is for more food for the reason that it means more food security. Africa needs wheat. Great strides are being made to get greater production of this product and greater food security in this region of the civilized world. The main objectives of this paper are as follows:

- To study previous trends in wheat production across different countries globally.
- To understand how wheat production changes over time and varies between countries.
- To develop a Machine Learning model that can forecast future wheat production accurately.
- To identify patterns and factors that influence wheat production.
- To provide useful information that can help government ministries, research and education councils, autonomous bodies, and international organizations that work to support farmers and promote agriculture make better decisions and improve overall wheat production.
- To support different countries in moving towards better food security and self-sufficiency in wheat production.

As of late, we have seen an increase in data available and, at the same time, improvements in data science, which have made machine learning (ML) a great tool for agricultural production prediction. Machine learning models are able to look at past data, identify trends that are out of sight to the naked eye, and do forecasting for what is to come [2]. Also, these models put at the disposal of governments, researchers, and farmers better tools, which in turn they use for resource allocation, supply chain management, and the development of agricultural policies. In our paper, we use a variety of machine learning regression models to predict crop production, which is our main focus on wheat production in African, Middle Eastern, and Oceanian countries. The data we looked at includes the year, country, crop type, production in terms of tonnes, domain, and year-over-year change. After cleaning and preprocessing of the dataset, these regression-based techniques are used:- Simple Linear-Regression [3], Multiple Linear-Regression, Polynomial Regression (Degree 2 - handled separately in the paper), Ridge Regression, Lasso Regression, Elastic Net Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression(SVR), KNN Regression, XGBoost Regression, Gradient Boosting Regression, AdaBoost Regression, CatBoost Regression, LightGBM Regression, etc. The regression models are evaluated and compared using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and R^2 Score to decide a suitable model for the prediction. By controlling overfitting, we shall try to obtain the best-performing model that can provide an accurate prediction on wheat production in countries globally.

1.1 Problem Statement

Wheat is one of the highly consumed food crops by humans in many African countries worldwide. It is an essential part of the food supply of vast numbers of people, and it provides a paying crop for many farmers. It is not, however, possible for Africa, the Middle East, and Oceania to produce sufficient quantities of wheat to feed its own population, and it is compelled to depend greatly on supplies from abroad. In fact, Africa produced in 2023 something like 28 million tons of wheat, which is far less than the quantity required. For this reason, countries are obliged to import wheat from other countries, which increases the costs to them and makes them subject to the fluctuations that control the prices in the general world market. It is obvious that this condition affects the question of food supply as well as the economic condition of the people. There are many reasons for the low and erratic production of wheat in countries globally [4]. Climatic difficulties such as irregularity of rainfall, drought, floods, and increased temperatures are being reflected directly in the growth and production of this important crop. The small farmer has usually not been provided with good-quality seed, fertilizer, or irrigation facilities. Because of this production of wheat is reduced. In addition to these many other ills, such as lack of government aid, lack of good marketing nets, etc., and poor agricultural equipment due to lack of development, militate against efficient wheat production. The older methods of forecasting wheat production have not been satisfactory, as they do not take into account the complicated relations between the different factors. However, existing studies primarily focus on region-specific datasets and fail to generalize across diverse agro-climatic zones such as Africa, the Middle East, and Oceania. These regions exhibit significant variability in climate patterns, soil conditions, and agricultural practices, which are often not adequately captured by traditional statistical or single-region machine learning models. Moreover, many existing approaches do not effectively model nonlinear dependencies or temporal variations in agricultural data. Therefore, there is a need for a generalized and robust predictive framework that can address these limitations and provide accurate cross-regional wheat production forecasts. This paper will attempt to correct this condition by applying the Machine Learning technique in forecasting wheat production in Africa. By taking into account the historical production figures along with climatic data, land use, etc., and other important factors, it is expected that fairly accurate results may be expected through the application of Machine Learning as an aid in these forecasts. In this modern world, Machine learning is helping in many fields, powering technologies and services that affect every decision of our daily lives. Machine learning helps in Everyday applications, Transportation and Autonomous Systems, Finance and Banking, healthcare, social media and Entertainment, manufacturing and industries, Education, etc. It can also help in the agricultural field with crop yield prediction, Disease Detection, Precision Farming, etc. These predictions will help government ministries, research and education councils, autonomous bodies, and international organizations that work to support farmers and promote agriculture make better decisions, improve wheat production, and ensure food security across the continent [5].

The related work of Wheat prediction using different techniques are shown in Table 1.

Table 1

Related work of wheat prediction using various machine, deep and quantum techniques

Author's Name & Year	Dataset	Methods Used for Analysis	Contribution	Research Results
Alhussan *et al.,* 2026 [6]	Multi-regional potato yield dataset	QTMs, ETNs, DPRNNs, STGCNs, preprocessing	First large-scale QTM application in agriculture	R^2=0.9998, lowest MSE
Nivethitha & Venkataramanan, 2025 [7]	Tomato & cotton leaf images	DTCWPF, EfficientNet-B7, ABQCNN, ASFA	Novel quantum-attention disease framework	Accuracy up to 99.92%
Bhanu *et al.,* 2025 [8]	Maize leaf IoT images	SA-MaxNet, AdSOA, FA-SegNet	Joint IoT-routing + DL framework	Accuracy 95.99%
Newman & Furbank, 2021 [9]	Australian NVT dataset	RF, XGBoost, SVM, satellite ML	Largest explainable ML agri study	R^2 > 0.8
Shahhosseini *et al.,* 2021 [10]	US Corn Belt APSIM + NASS	APSIM + ML hybrids	Systematic hybrid evaluation	RMSE reduced 7–20%
Cvejić *et al.,* 2023 [11]	Sunflower hybrid dataset (1250 samples)	ANN, SVR, KNN, Random Forest	First ML-based sunflower oil yield study	RFR achieved R^2=0.92
Abate *et al.,* 2024 [12]	Satellite + historical yield data	RF, Gradient Boosting, KNN, Decision Trees	Region-specific ML framework	RF outperformed others
Ashfaq *et al.,* 2025 [13]	Satellite, climate, soil data (2017–2022)	CNN, RNN, ANN (DeepAgroNet)	Novel multi-branch DL model	CNN R^2=0.77
Alsaber *et al.,* 2025 [14]	16-year weather & yield data	ANN, RF, XGBoost, SVR, ELNET	District-wise ML comparison	ANN best (>98% accuracy)
Setiadi *et al.,* 2024 [15]	Rice production datasets	Quantum circuits + Bi-LSTM + XGBoost	First hybrid quantum DL rice study	R^2=0.9999
Gautam *et al.,* 2025 [16]	NASA POWER, USDA, NASS (2000–2024)	AquaCrop, Semi-physical, ANN	Comparative ANN vs crop models	ANN R^2=0.96
Ravikumar *et al.,* 2025 [17]	Tamil Nadu crop reports	LSTM, SVR, XGBoost ensemble	Ensemble ML literature	R^2=0.9918
Tamayo-Vera *et al.,* 2025 [18]	Canada PEI (1982–2016)	RF, Gradient Boosting, SHAP	ML yield studies	RF RMSE lowest
Banadkooki *et al.,* 2026 [19]	Iran groundwater data	PSO-COO-GRU-ANFIS	AI hydrology studies	NSE=0.90
Yan *et al.,* 2025 [20]	Literature review	ML, DL, sensors review	Extensive survey	Trends summarized
Regaieg *et al.,* 2025 [21]	In-situ and airborne SIF measurements (alfalfa crop)	DART radiative transfer model + Fluspect-CX	Novel physical SIF downscaling framework	R^2 up to 0.76 with PSII measurements
Klinnert *et al.,* 2025 [22]	FAOSTAT, climate, soil data	DSSAT crop model + XGBoost	Large-scale ML–crop model integration	Yield increase 14–17%
Jhajharia *et al.,* 2023 [23]	Historical crop & weather data	RF, SVM, LSTM, Gradient Descent	Comparative ML evaluation for Indian crops	RF achieved R^2 = 0.963
Al-Shamayleh & Ibrahim, 2025 [24]	Kaggle grape disease dataset	(q,τ)-Nabla calculus + CNN features	Novel quantum-deformation feature fusion	Accuracy improved over baseline
Cui *et al.,* 2026 [25]	108 PFAS experiments + 6203 predicted PFASs	Molecular docking + ML	First crop-specific AOP framework	High-risk PFAS identified

2. Methodology

2.1 Problem Definition

The main objective is to predict wheat yield (in tonnes) based on country and year. The input features include year, country, flag (estimated or official), and YOY Change, while the output variable represents the crop yield in tonnes that the country produces. The workflow diagram is shown in Figure 1.

2.2 Data collection

The dataset "global-wheat.csv" was obtained from Kaggle, which is a merged dataset of 3 datasets:

1. Africa Wheat Production Data (1961–2025)
2. Middle East Wheat Production Data (1961–2025)
3. Oceania Region Wheat Production Dataset

This merged dataset has a total of 1300 rows with 9 columns. The rows hold data on the annual yield of wheat production, the country name, year, flag (actual or predicted), and the year-over-year (YoY) indices, which indicate the upcoming agricultural performances and food supply superiority indices that prevail in each country from Africa, the Middle East, and Oceania from 1961-2025.

This data gives a comprehensive review of the wheat production trends existing in 19 countries (Algeria, Morocco, Ghana, Ethiopia, Kenya, Egypt, South Africa, Nigeria, Samoa, Papua New Guinea, Fiji, New Zealand, Australia, Jordan, Israel, Iraq, Turkey, Iran, and Saudi Arabia) across Africa, Middle East, and Oceania from 1961 to 2025.

2.3 Data description

Year - Year of production data
Value - Quantity of Wheat produced
Unit - Value unit (Here in tonnes)
Flag - Estimated or Official
Country - Area of Data Collected
Item - Name of Crop (Wheat)
Domain - Data domain or sector (e.g., Production)
Metric - The type of metric recorded (e.g., Production)
YoY Change - Year-over-year change (in Percentage)

2.4 Data preprocessing

The dataset was loaded in the form of a data frame using the read_csv() function. The data frame contained a total of 65 duplicated values. These duplicated values were removed before proceeding further using the drop_duplicated method. After removing the duplicated rows, the data frame now has a shape of (1235 rows × 9 columns). In the data set, about 9 values were missing from the column YoY Change. These missing values were replaced by the mean of the 'YoY Change' feature of the dataset. The categorical features, like Country, Item, Domain, Metric, Flag, and Unit, were encoded using LabelEncoder, which comes with the Scikit-Learn library. Features determining the output were selected using feature selection. And the dataset was divided into input(X) and output(y). After this step, input and output were further split into training data and testing data using train_test_split() in the ratio 80:20. From this step, we get X_train, which will be used for training the model, and X_test, which will be used for evaluation of the model performance. These training and testing data were scaled using StandardScaler provided by the Scikit-Learn library. In addition to handling missing values and duplicates, outlier detection was performed using the interquartile range (IQR) method to identify extreme values that could adversely affect model performance. Any detected outliers were carefully examined and retained or removed based on domain relevance. Furthermore, assumptions such as data stationarity and consistency across years were considered during preprocessing to ensure reliability of the predictive models.

2.5 Model selection and training

The paper evaluation included various regression methods including: Simple Linear-Regression, Multiple Linear-Regression, Polynomial Regression (Degree 2 - reviewed differently), Ridge Regression, Lasso Regression, Elastic Net Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression (SVR), KNN Regression, XGBoost Regression, Gradient Boosting Regression, AdaBoost Regression, CatBoost Regression, and LightGBM Regression. The selection of ensemble-based models such as XGBoost and LightGBM was motivated by their ability to handle complex nonlinear relationships and interactions among features, which are common in agricultural datasets. Compared to traditional regression models, these algorithms provide better generalization and robustness against overfitting. Additionally, boosting-based models iteratively improve prediction performance by minimizing residual errors, making them particularly suitable for time-dependent and climate-sensitive data. Max_depth and n_estimator of Decision trees or Random Forests were set to a maximum of 6 and a maximum of 200 respectively to avoid overfitting while tuning the learning rate for some boosting algorithms. The best performing method for robustness and non-linearity was selected for XGBoost. Cross-Validation of 5-folds ewas utilized due to reducing the risk of overfitting. Despite their advantages, the selected machine learning models have certain limitations. For instance, linear models may struggle with nonlinear dependencies, while tree-based ensemble models can be sensitive to multicollinearity among features. Additionally, these models rely heavily on historical data and may not fully capture sudden environmental changes or extreme climatic events.

2.6 Model evaluation

The performance of the ML models was assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score and XGBoost achieved the best R^2 score of 0.98.

2.7 Tools and technologies

This paper is implemented using Python and its libraries such as NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, XGBoost, CatBoost, and LightGBM. The paper was developed using Jupyter Notebook, which is a tool that comes with Anaconda Navigator.

3. Experiment

The objective of this study is to predict the Wheat Yield across different countries of the African continents using various regression algorithms and compare their performance. The goal is to determine which model provides the highest prediction accuracy and stability over time.

3.1 Experimental Result Setup

The experiments were carried out on a system equipped with an NVIDIA RTX 3050 GPU, 16 GB RAM, and an Intel Core i5 processor. The development environment consisted of the following technologies:

- Programming Language: Python 3.10
- Libraries used:
  - For data handling: NumPy, Pandas
  - For data visualization: Matplotlib, Seaborn
  - Machine Learning Models: XGBoost, LightGBM, CatBoost etc.
  - Evaluation: scikit-learn.metrics

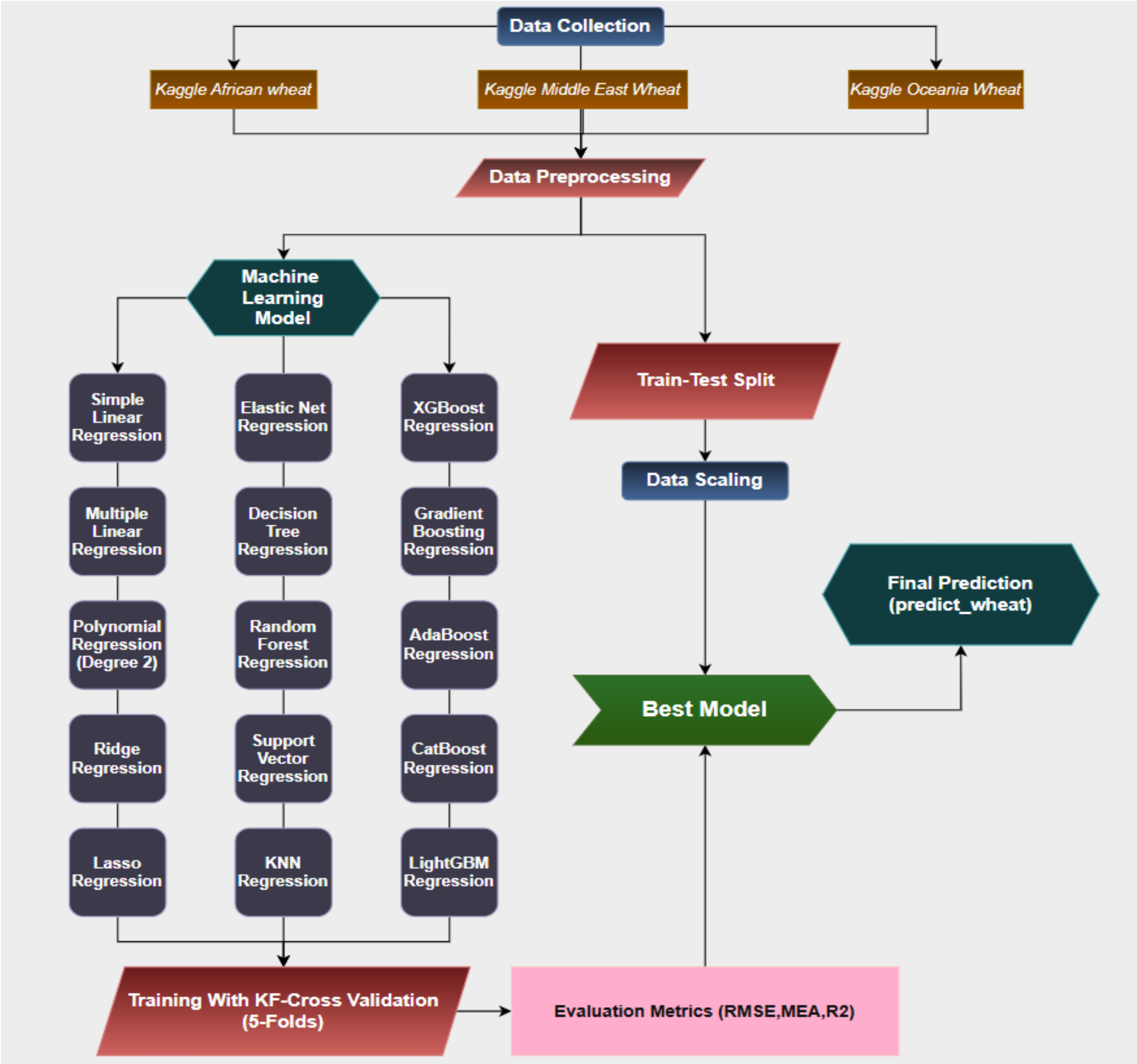

Fig. 1. Proposed framework

3.2 Importing Necessary Libraries and DataFrame Creation

In this paper we import libraries like NumPy, Pandas, Scikit-learn, MatplotlibPy, Seaborn, XGBoost, CatBoost, LightGBM and used their Regression training Methods. Using Pandas, we prepare a DataFrame from the provided dataset “africa_final - africa.csv”, having the Shape (1300, 9) i.e., 1300 rows and 9 columns. The columns are: Year, Value, Unit, Flag, Country, Item, Domain, Metric, YoY Change.

3.3 Feature Selection and Data Cleaning

The dataset used for this experiment contains records of annual wheat production in various African countries. The attributes selected include Year, Country, Yield (in tonnes), and Year-over-Year (YoY) change percentage as shown in Table 2. Before training, the data was carefully examined to remove inconsistencies and missing entries were replaced with mean value of that column. To ensure that all input variables contributed equally, numerical values were normalized. The information and data type of each feature in the dataset is given in Table 3.

Table 2

Sample of features

Year	Value	Unit	Flag	Country	Item	Domain	Metric	YoY Change
1961	1000000	tonnes	Estimated	Algeria	Wheat	Production	Production	
1962	1020000	tonnes	Official	Algeria	Wheat	Production	Production	2
1963	1040000	tonnes	Estimated	Algeria	Wheat	Production	Production	1.960784
1964	1060000	tonnes	Official	Algeria	Wheat	Production	Production	1.923077
1965	1080000	tonnes	Estimated	Algeria	Wheat	Production	Production	1.886792
1966	1100000	tonnes	Official	Algeria	Wheat	Production	Production	1.851852
1967	1120000	tonnes	Estimated	Algeria	Wheat	Production	Production	1.818182
1968	1140000	tonnes	Official	Algeria	Wheat	Production	Production	1.785714
1969	1160000	tonnes	Estimated	Algeria	Wheat	Production	Production	1.754386
1970	1180000	tonnes	Official	Algeria	Wheat	Production	Production	1.724138

Table 3

Data types of features

Column	Non-Null Count	Dtype
Year	1300 non-null	int64
Value	1300 non-null	int64
Unit	1300 non-null	object
Flag	1300 non-null	object
Country	1300 non-null	object
Item	1300 non-null	object
Domain	1300 non-null	object
Metric	1300 non-null	object
YoY Change	1280 non-null	float64

3.4 Encoding Categorical Features

In this step, the categorical features Country, Item, Domain, Metric, Flag, and Unit were transformed into numerical values using Label Encoding.
A basic feature encoder (LabelEncoder) was used on each column to provide a unique integral code to their distinct categories. For example, in flag column the LabelEncoder gave ‘0’ code for estimated entries and ‘1’ for official entries. This conversion ensures that the categorical data can be efficiently considered and utilized by machine learning algorithms during model training and evaluation.

3.5 Visualising the data correlation

A correlation heatmap between Year, Value (Wheat Yield), Country & YoY Change was created to visually represent the relationships between the previously mentioned numerical and encoded data elements. The heatmap provides an easy-to-see visual representation of how well the different features are linked together with respect to one another; as well as showing the strength of the relationship between the features using color gradation.

3.6 Model implementation

All of the abovementioned experiment(s) were performed using several different Python Libraries within the Jupyter Notebook/JupyterLab Environment. The machine learning library that was used in this scenario is Scikit-learn, XGBoost, LightGBM and CatBoost along with Supporting Libraries for statistical analysis such as NumPy as well as Matplotlib and Seaborn for visual representation of our data. The regression models that were provided that assess which was the best model for conducting prediction analysis have been included Simple Linear Regression, Multiple Linear Regression, Ridge Regression, Lasso Regression etc.

3.6.1 Polynomial regression (Degree 2)

Polynomial Regression is a form of linear regression where the relationship between the independent variable (x) and the dependent variable (y) is modelled as an n^{th} degree polynomial.

Formula (Degree 2) –

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Formula (In general) –

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n + \varepsilon$$

3.6.2 Elastic net regression

Elastic net regression combines L1 (Lasso) and L2 (Ridge) penalties as forms of regularization with linear regression, providing both feature selection and shrinkage. Elastic net regression is an excellent option when you have many variables to work with and they are correlated because elastic net groups correlated variables together, allowing for their removal as one entity or individual retainment of any respective correlations.

3.6.3 Random Forest regression

Random Forest is bagging algorithm that uses many weak learner decision trees to make better predictions. Random Forest is a like a team of Decision Trees that work together on different subset of the dataset and the subset datasets are created using Bootstrap sampling (Row subsets) and Random Feature selection (Column selection) using different sampling methods. The result is combined by voting for classification problems like fraud detection, disease diagnosis etc, and averaging for regression problems like price prediction, weather forecasting, crop yield prediction etc.

3.6.4 Support vector regression

Support Vectors are points or groups that are located closest to a hyperplane, a line (in 2 dimensions) or a plane (in 3 dimensions) which separates two different classes. Support Vector Regression (SVR) uses kernels, generates a sparse solution, and finds a sufficient number of support vectors while controlling for margin. SVR is less common than SVM (as a function of its effectiveness) but has proven to be a very useful tool for estimating real value functions. As a supervised Learning method, SVR has a symmetric loss function used during training, which treats high and low estimates equally.

4.6.5 KNN regression

The application of the KNN algorithm (like other Machine Learning algorithms) is not dependent on prior assumptions about the functional relationship between the dependent and independent variables. KNN regression works on predicting a continuous output for an observation. In KNN regression, to predict the target value of the new observation, we take the mean of the target values of the K nearest observations (neighbours) in feature space. Typically, distance between observations is computed using Euclidean distance; however, other distance metrics may also be employed for this computation.

4.6.6 XGBoost regression

Extreme Gradient Boosting is an optimized and scalable version of Gradient Boosting, which is used for supervised learning tasks (e.g., classification, regression). The goal of XGBoost for regression is to predict a numeric value (i.e., continuous) using an iterative training process that minimizes loss function (i.e., RMSE or MSE) with the use of regularization techniques to prevent overfitting.

4.6.7 Gradient boosting regression

Gradient Boosting is a strong and adopted machine learning methodology for solving both regression and classification tasks. This framework creates its models in a stepwise fashion whereby new models learn from their predecessors and improve predictions based on those errors. During each iteration of the training process, the Gradient Boosting algorithm determines the direction and magnitude of change (the gradient) of the loss function as a result of the current predictions and fits an additional weak model which is capable of reducing this value further. In later iterations, predictions from newly created models will also contribute to the overall predictions of the ensemble of models trained thus far. The algorithms will continue to improve creating new models until they have reached some acceptable performance level (stopping conditions).

4. Results and Discussions

4.1 Model Configuration with Parameter Tuning

For every algorithm, parameters were tuned to obtain optimal results. Hyperparameter tuning was performed using a grid search strategy combined with cross-validation to identify optimal parameter values for each model. The selection of parameters such as learning rate, maximum depth, and number of estimators was based on minimizing prediction error while preventing overfitting. The chosen values represent a balance between model complexity and generalization performance. The parameter tuned for each model are given below:

‘Ridge Regression: alpha=1.0’,’Lasso Regression: alpha=0.1’,
‘Elastic Net Regression: alpha=0.1’, ‘l1_ratio=0.5,
Decision Tree Regression: ‘max_depth’=6, ‘min_samples_split’=10,
Random Forest Regression: ‘n_estimators’=200, ‘max_depth’=6, ‘random_state’=42,
Support Vector Regression": ‘C’=0.01, ‘epsilon’=0.1, kernel='rbf',
KNN Regression: ‘n_neighbors’=5,
XGBoost Regression: ‘n_estimators’=200, ‘max_depth’=4, ‘learning_rate’=0.05,
Gradient Boosting Regression: ‘n_estimators’=200, ‘max_depth’=5, ‘learning_rate’=0.05,
AdaBoost Regression : ‘n_estimators’=100, ‘learning_rate’=0.1,
CatBoost Regression: ‘verbose’=0, ‘depth’=5, ‘learning_rate’=0.05, ‘iterations’=200,
LightGBM Regression: ‘max_depth’=5, ‘learning_rate’=0.05, ‘n_estimators’=200

By visually displaying how well the features relate to one another with the correlation heatmap shown in Figure 2, it allows us to identify patterns and dependencies between the independent variables and informs future data analysis and model development.

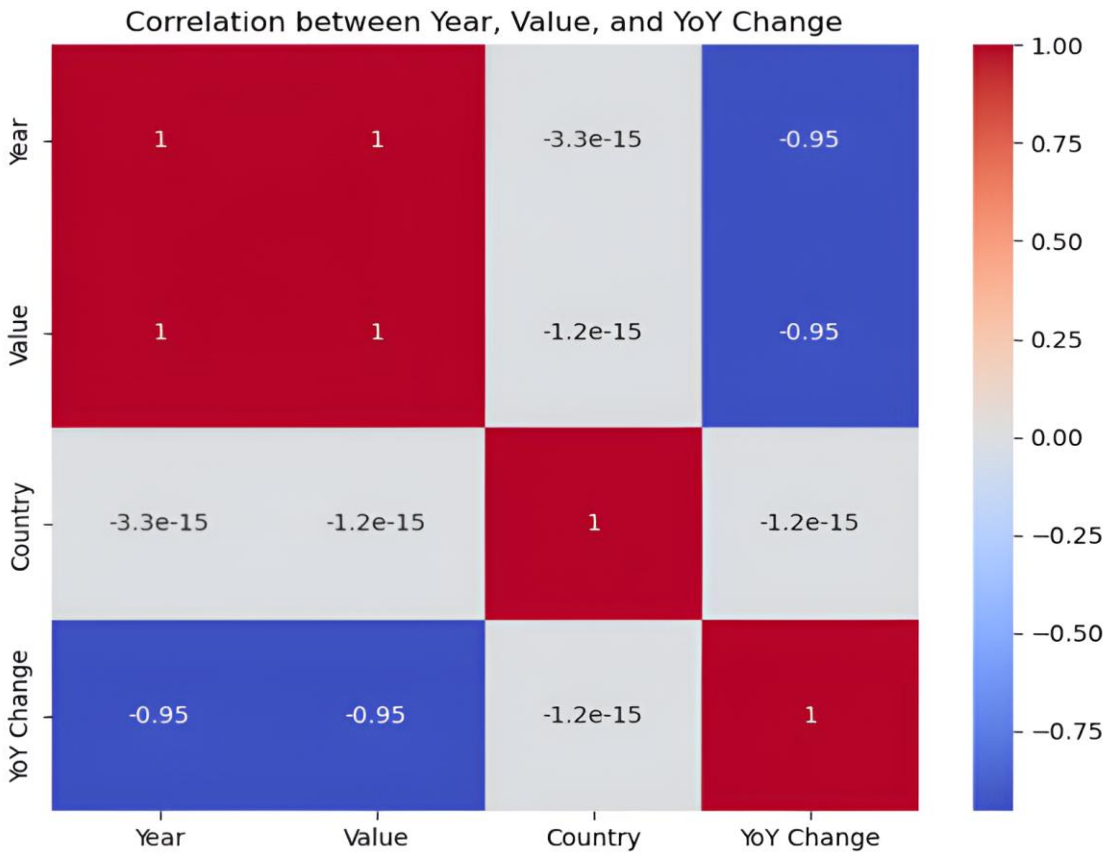

Fig. 2. Confusion matrix

4.2 Evaluation Metrics with Cross Validation

Three major performance evaluation metrics, MAE, RMSE and R^2, were utilized to conduct the overall evaluation of the models. RMSE was selected as it penalizes larger errors more significantly, making it suitable for evaluating prediction reliability in agricultural forecasting. MAE provides an intuitive measure of average prediction error, while R^2 indicates the proportion of variance explained by the model. Together, these metrics offer a comprehensive evaluation of model performance in real-world agricultural scenarios. Each of these metrics shows how well each of the models predicted yield from the actual yield and how well each model will generalize to the yield of future samples shown in Table 4 and Figure 3.

The superior performance of XGBoost can be attributed to its ability to model complex nonlinear interactions between climatic and temporal variables. The results indicate that features such as year-over-year change and country-specific patterns significantly influence wheat production. Climatic variability, particularly rainfall and temperature fluctuations, plays a critical role in determining yield outcomes, as reflected in the model predictions.

Intelligent Systems Research and Applications Journal
Volume 2, (2026) 288-304

Table 4

Results of ML models

Model	MAE	RMSE	R^2 Score
XGBoost Regression	13.9622	18.3707	0.999999
Polynomial Regression (Degree 2)	0.1236	0.240436	0.983745
Simple Linear Regression	0.133107	0.2456	0.999988
Multiple Linear Regression	0.1331	0.2454	0.999969
Lasso Regression	1.4262	2.100940	0.999995
Gradient Boosting Regression	11.056	13.0830	0.999858
Random Forest Regression	252.77	362.44	0.999999
LightGBM Regression	252.60	853.15	0.999995
Ridge Regression	873.20	1289.33	0.999988
Decision Tree Regression	347.394541	2063.585398	0.999969
CatBoost Regression	3387.616480	4393.969986	0.999858
KNN Regression	10947.394332	14124.999403	0.998531
Elastic Net Regression	21749.514045	31930.298359	0.992493
AdaBoost Regression	41827.478493	46984.703000	0.983745
Support Vector Regression	313764.039439	368835.450557	-0.001718

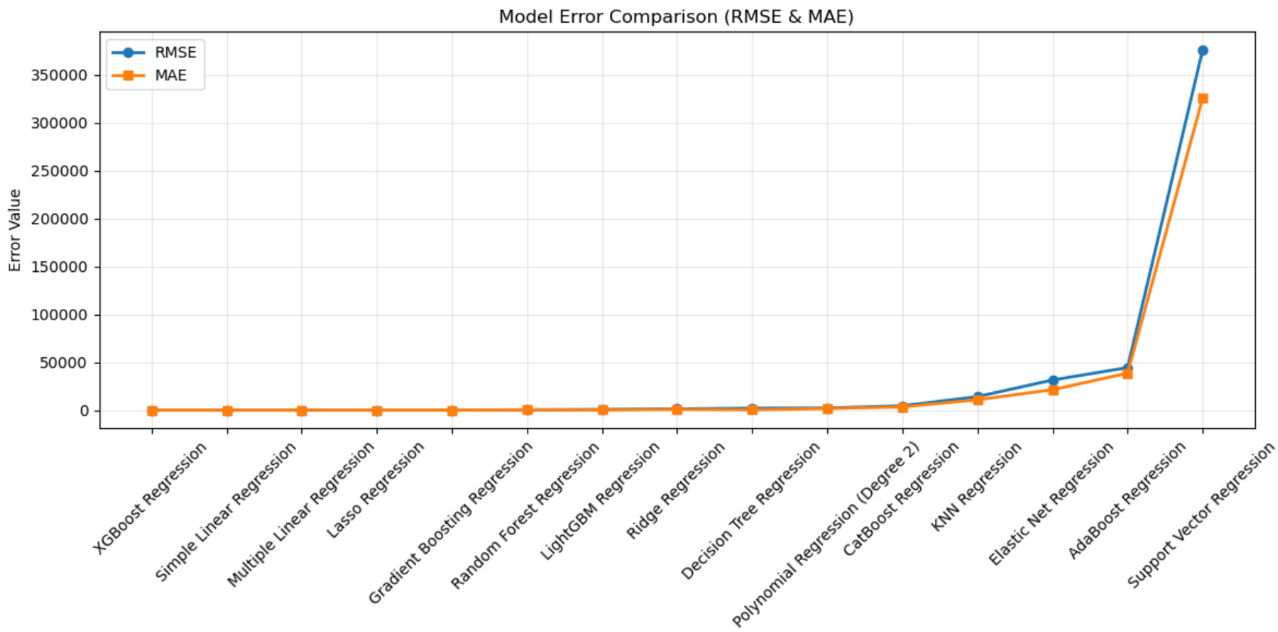

Fig. 3. RMSE and MAE comparison of ML models

Research indicates that machine learning is an effective method for predicting wheat yield in areas with complex combinations of environmental and economic conditions. Three machine learning algorithms, Random Forest, XGBoost, and CatBoost were identified to have the most promise using their ability to identify nonlinear relationships among inputs and establish associations from multiple perspectives. Through analyzing the variable importance associated with wheat yield, it was revealed that climate-related variables (i.e., rainfall and temperature variability) were the largest contributors to wheat yield across all three geographical regions assessed. It was also determined through analysis of the three regions that environmental factors affect the amount of crop produced in different ways between geographical areas, supporting the development of agriculture specific to a particular area. Through the use of predictive analytics, this paper successfully provides information and guidance to help in the decision-making process of policymakers, farmers, and researchers regarding the allocation of resources, the provision of food security and the establishment of sustainable agriculture practices. As a result of this work, a strong foundation has been established to support future advancements, including the addition of real-time weather data to the model, expanding the geographical area studied, and developing the predictive model into a web-based tool that can support decision-making. The Best model that performed on the dataset is XGBoost Regressor with 0.99 R^2 Score, and the function "predict_wheat()" is used to predict the wheat production which takes attributes Country of *string* type, Year of *integer* type, Flag that tells whether to predict *Estimated value, or Official Value*, and approximate YoY change in *float.* For instance, policymakers can utilize the predictive outputs to optimize import-export strategies, allocate resources for irrigation infrastructure, and implement region-specific agricultural policies. In drought-prone regions, early predictions can support contingency planning, while in high-yield regions, surplus management strategies can be developed.

5. Conclusions

This research focuses on developing a machine learning system capable of predicting wheat production across three diverse regions Africa, the Middle East, and Oceania. The paper began with the collection and preprocessing of historical data, such as Year of record, Wheat yield that year (in tonnes), flag i.e., whether the record is official or estimated, Country of Record and Y-o-Y Change in yield. In order to provide a high-quality, consistent dataset, considerable effort was made to clean up the historical datasets through ways like normalization, outlier detection and the conversion of categorical variables into numerical representations of categorical variables to allow them to be put into machine learning algorithms. Machine learning techniques that were adopted include Linear Regression, Random Forest, Logistic Regression, XGBoost, LightGBM and CatBoost to determine which algorithm produced the most consistently accurate and robust predictions. To further improve the models, several techniques were used, such as cross-validation, hyperparameter tuning, feature importance analysis, and the visualization of data trends. With these evaluations, we were able to fairly evaluate the performance of each of the models with the help of RMSE, MAE and R^2 score metrics. The most accurate performing model was subsequently integrated into a wheat yield estimation pipeline to provide regional wheat yield predictions. The complete end-to-end workflow demonstrates how to handle machine-learning from data acquisition through to final product ionisation of predictions for a region using a machine-learning-based solution. Future work may focus on integrating real-time weather data, satellite imagery, and soil health indicators to enhance prediction accuracy. Additionally, extending the framework to other crops and geographical regions would improve its generalizability. The incorporation of deep learning and hybrid models could further enhance performance in capturing temporal and spatial dependencies.

Author Contributions

Authors individual contributions must be provided. The following statements should be used "Conceptualization, Anju,Naman, Paras Dahiya; methodology Paras Dahiya, Naman; software, Anju, Naman, Suraj Arya, Paras Dahiya; validation, Anju., Naman. and Suraj Arya.; formal analysis, Anju.; investigation, Suraj Arya.; resources, Suraj Arya, Anju. and Naman; data curation, Anju; writing—original draft preparation, Anju. Naman. Suraj Arya.; writing—review and editing, Paras Dahiya Anju. Naman. Suraj. Arya.; visualization, Anju. Naman. Suraj Arya.; supervision, Suraj Arya.; project administration, Suraj Arya and Anju.; All authors have read and agreed to the published version of the manuscript. Authorship must be limited to those who have contributed substantially to the work reported.

Funding

This research received no external funding.

Data Availability Statement

Available on request.

Conflicts of Interest

No conflicts among authors.

Acknowledgement

Authors thanks to the Central University of Haryana for proving valuable resources to conduct this research

References

[1] Ahmed, M. U., & Hussain, I. (2022). Prediction of wheat production using machine learning algorithms in northern areas of Pakistan. *Telecommunications policy*, *46*(6), 102370. https://doi.org/10.1016/j.telpol.2022.102370

[2] Iqbal, N., Shahzad, M. U., Sherif, E. S. M., Tariq, M. U., Rashid, J., Le, T. V., & Ghani, A. (2024). Analysis of wheat-yield prediction using machine learning models under climate change scenarios. *Sustainability*, *16*(16), 6976. https://doi.org/10.3390/su16166976

[3] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69-134). Cham: Springer international publishing. https://doi.org/10.1007/978-3-031-38747-0_3

[4] Haseeb, M., Tahir, Z., Mahmood, S. A., & Tariq, A. (2025). Winter wheat yield prediction using linear and nonlinear machine learning algorithms based on climatological and remote sensing data. *Information Processing in Agriculture*. https://doi.org/10.1016/j.inpa.2025.02.004

[5] Awad, M., & Khanna, R. (2015). Support vector regression. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 67-80). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

[6] Ali Alhussan, A., Radwan, M., Khodadadi, N., Mattar, E. A., & Kenawy, S. (2026). Towards Intelligent Agricultural Forecasting: Quantum Temporal Models for Potato Yield Prediction. *Potato Research*, 69(2), 36. https://doi.org/10.1007/s11540-025-09971-w

[7] Nivethitha, T., & Venkataramanan, C. (2025). A Quantum-Enhanced Deep Learning Framework for Precise Leaf Disease Detection in Tomato and Cotton Crops. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 1-32. https://doi.org/10.1007/s40998-025-00912-4

[8] Bhanu, K. S., Gunisetti, L., Vinnakota, S., Jagannathan, V., & Ramesh, C. R. (2026). Maize Leaf Disease Classification with Shuffle Attention Max Network and Optimized Internet of Things Routing. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 1-26. https://doi.org/10.1007/s40998-025-00980-6

[9] Newman, S. J., & Furbank, R. T. (2021). Explainable machine learning models of major crop traits from satellite monitored continent-wide field trial data. *Nature Plants*, *7*(10), 1354-1363. https://doi.org/10.1038/s41477-021-01001-0

[10] Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific reports*, 11(1), 1606. https://doi.org/10.1038/s41598-020-80820-1

[11] Cvejić, S., Hrnjaković, O., Jocković, M., Kupusinac, A., Doroslovački, K., Gvozdenac, S., ... & Miladinović, D. (2023). Oil yield prediction for sunflower hybrid selection using different machine learning algorithms. *Scientific Reports*, 13(1), 17611. https://doi.org/10.1038/s41598-023-44999-3

[12] Abate, J., Aman, A., & Adem, D. (2025). Integration of satellite data for predicting crop yields in Eastern Ethiopia using machine learning. *Scientific Reports*, 15(1), 33809. https://doi.org/10.1038/s41598-025-00810-z

[13] Ashfaq, M., Khan, I., Shah, D., Ali, S., & Tahir, M. (2025). Predicting wheat yield using deep learning and multi source environmental data. *Scientific Reports*, 15(1), 26446. https://doi.org/10.1038/s41598-025-11780-7

[14] Alsaber, A., Satpathi, A., Alsabah, M., & Setiya, P. (2025). Optimizing potato yield predictions in Uttar Pradesh, India: a comparative analysis of machine learning models. *Scientific Reports*, 15(1), 26897. https://doi.org/10.1038/s41598-025-12719-8

[15] Setiadi, D. R. I. M., Susanto, A., Nugroho, K., Muslikh, A. R., Ojugo, A. A., & Gan, H. S. (2024). Rice yield forecasting using hybrid quantum deep learning model. *Computers*, *13*(8), 191. https://doi.org/10.3390/computers13080191

[16] Gautam, V., Gani, A., Pathak, S., & Shukla, A. K. (2025). Evaluating crop yield prediction models in illinois using aquacrop, semi-physical model and artificial neural networks. *Scientific Reports*, 15(1), 27494. https://doi.org/10.1038/s41598-025-13453-x

[17] Ravikumar, A., Periyasamy, V., Devi, K. M. K., Krishnamoorthy, R. G., & Mariavalavan, O. K. (2025). A hybrid SERWI ensemble model for crop yield prediction using an inverse RMSE weighting strategy. *Scientific Reports*. https://doi.org/10.1038/s41598-025-31987-y

[18] Tamayo-Vera, D., Mesbah, M., Zhang, Y., & Wang, X. (2025). Advanced machine learning for regional potato yield prediction: analysis of essential drivers. *npj Sustainable Agriculture*, *3*(1), 12. https://doi.org/10.1038/s44264-025-00052-6

[19] Banadkooki, F. B., Ghanbari-Adivi, E., Sayyahi, F., & Ehteram, M. (2026). An intelligent hybrid deep learning machine learning model for monthly groundwater level prediction. *Scientific Reports*. https://doi.org/10.1038/s41598-025-34292-w

[20] Yan, J., Wu, H., Diao, Z., Miao, Y., Zhang, B., & Zhao, C. (2025). Recent Developments and Applications of Crop Disease Detection, Prediction, and Early Warning: A Review. *Engineering*. https://doi.org/10.1016/j.eng.2025.10.032

[21] Regaieg, O., Malenovský, Z., Siegmann, B., Buffat, J., Krämer, J., Lauret, N., & Le Dantec, V. (2025). DART-based temporal and spatial retrievals of solar-induced chlorophyll fluorescence quantum efficiency from in-situ and airborne crop observations. *Remote Sensing of Environment*, 319, 114636. https://doi.org/10.1016/j.rse.2025.114636

[22] Klinnert, A., Rogna, M., Barbosa, A. L., Tillie, P., & Baldoni, E. (2025). The potential of irrigation for cereals production in Sub–Saharan Africa: A machine learning application for emulating crop growth at large scale. *Agricultural Water Management*, *314*, 109488.https://doi.org/10.1016/j.agwat.2025.109488

[23] Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop yield prediction using machine learning and deep earning techniques. *Procedia Computer Science*, *218*, 406-417. https://doi.org/10.1016/j.procs.2023.01.023

[24] Al-Shamayleh, A. S., & Ibrahim, R. W. (2025). Grapevine disease detection using (q, τ)-nabla calculus quantum deformation with deep learning features. *MethodsX*, 103619. https://doi.org/10.1016/j.mex.2025.103619

[25] Cui, Y., Xiao, W., Fang, Y., Li, B., Zheng, M., & Li, Y. (2026). PFAS bioaccumulation in food crops: Structure informed machine learning and mechanistic evidence for priority control. *Environmental Chemistry and Ecotoxicology*. https://doi.org/10.1016/j.enceco.2025.12.029