IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 January 2024, accepted 20 March 2024, date of publication 29 March 2024, date of current version 5 April 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3383309*

RESEARCH ARTICLE

# Incorporating Meteorological Data and Pesticide Information to Forecast Crop Yields Using Machine Learning

**MD JIABUL HOQUE[1,2], MD. SAIFUL ISLAM[2], JIA UDDIN[3], MD. ABDUS SAMAD[4], (Member, IEEE), BEATRIZ SAINZ DE ABAJO[5], DÉBORA LIBERTAD RAMÍREZ VARGAS[6,7,8], AND IMRAN ASHRAF[4]**

[1]Department of Computer and Communication Engineering, International Islamic University Chittagong, Kumira, Chattogram 4318, Bangladesh
[2]Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering and Technology, Chittagong 4349, Bangladesh
[3]AI and Big Data Department, Endicott College, Woosong University, Daejeon 34606, South Korea
[4]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si 38541, South Korea
[5]Department of Signal Theory, Communications and Telematics Engineering, University of Valladolid, 47011 Valladolid, Spain
[6]Universidad Europea del Atlántico, 39011 Santander, Spain
[7]Universidad Internacional Iberoamericana, Campeche 24560, Mexico
[8]Universidad de La Romana, La Romana, Dominica

Corresponding authors: Md. Saiful Islam (saiful05eee@cuet.ac.bd), Md. Abdus Samad (masamad@yu.ac.kr), and Imran Ashraf (ashrafimran@live.com)

This work was supported by European University of Atlantic.

**ABSTRACT** The agricultural sector is more vulnerable to the adverse effects of climate change and excessive pesticide application, posing a significant risk to global food security. Accurately predicting crop yields is essential for mitigating these risks and providing information on sustainable agricultural practices. This research presents a novel crop yield prediction system that utilizes a year's worth of meteorological data, pesticide records, crop yield data, and machine learning techniques. We employed rigorous methods to gather, clean, and enhance data and then trained and evaluated three machine learning models: Gradient Boosting, K-Nearest Neighbors, and Multivariate Logistic Regression. We utilized the GridSearchCV method for hyper-parameter tweaking to identify the most suitable hyper-parameter throughout K-Fold cross-validation, aiming to improve the model's performance by avoiding overfitting. The remarkable performance of the Gradient Boosting model, with an almost flawless coefficient of determination ($R^2$) of 99.99%, demonstrates its promise for precise yield prediction. This research also examined the correlation between projected and actual crop yields and identified the ideal meteorological conditions. It paves the way for data-driven methods in sustainable agriculture and resource distribution, ultimately leading to a more secure future with respect to food availability and resilience to climate change.

**INDEX TERMS** Agriculture, crop yield prediction, machine learning, deep learning.

## I. INTRODUCTION

Aagriculture is an economic endeavor that is significantly dependent on meteorological conditions [1]. The viability of seasonal agriculture depends on the prevailing natural weather conditions, sometimes called rainfed agriculture. Rainfed agriculture, which covers approximately 80% of global cropland, demonstrates favorable crop yields when the weather conditions are favorable [2]. It is essential to recognize that agricultural productivity continues to be significantly dependent on precipitation and several meteorological factors [3]. In certain cases, farmers may need more time to obtain the anticipated crop yield due to variations in rainfall and other meteorological factors, either due to scarcity or

The associate editor coordinating the review of this manuscript and approving it for publication was Liandong Zhu.

excess [4]. Based on the data mentioned earlier, the prediction of crop production poses a significant challenge within the field of precision agriculture [5].

Climate change has a high impact on the agricultural sector, which can lead to adverse outcomes such as food poverty and famine [6]. Precipitation and temperature are pivotal climatic factors impacting agricultural production, influencing secondary factors such as soil moisture and solar irradiance. Focusing research on key variables offers valuable insights into crop yields, streamlining efforts for effective strategies to safeguard food security amid a changing climate [7]. Multiple research investigations have demonstrated that climatic indicators, whether on a global or regional scale, significantly impact agricultural yields and the overall state of food security [3], [5], [6], [8], [9]. A recent study conducted by Javadinejad et al. [10] discovered a correlation between diminished crop yields and two environmental factors: elevated temperatures and increased precipitation. We can attribute the adverse effects of severe temperatures on agricultural production to several factors, including increased rates of evapotranspiration and respiration in crops and a heightened vulnerability to pest infestation. An augmentation in the intensity of precipitation leads to amplified flow patterns, thus resulting in the occurrence of floods and heightened vulnerability to crop failure. The increase in temperature can affect crop productivity due to the subsequent increase in the demand for water for crops [11]. Although climate factors may remain consistent within a particular area, it is essential to note that the requirements of the weather parameters vary between different crops according to their respective growth stages [12]. Each crop exhibits varying degrees of resilience in response to meteorological conditions. When meteorological factors increase significantly to an extreme degree, there will be a notable impact on crop yield [13].

The agricultural sector in India is highly dependent on the monsoon season [14], [15]. In India, around 70% of the annual precipitation occurs during the monsoon season, facilitating irrigation for approximately half of the country's agricultural fields. The available data suggest that fluctuations in monsoon rainfall have the potential to result in significant changes in crop production. A deficient monsoon in 2014 resulted in a reduction of around 5% in total crop production [16], [17], [18]. Insufficient precipitation and drought phenomena have profound implications. Based on an analysis of historical data, we have observed that drought can decrease crop yields from 7% to 10% [19]. Excessive precipitation, frequently linked to cyclonic activity and inundations, can potentially result in waterlogging and soil erosion [20]. Intense monsoon rains and subsequent floods in 2019 led to substantial agricultural yield reductions, notably in regions such as Maharashtra and Karnataka [21]. Elevated temperatures experienced during crucial growth phases have the potential to induce heat stress, hence diminishing crop yields. The research findings indicate a negative correlation between temperature increase and crop yields during the sensitive blooming stage of the crop, with a reduction of approximately 0.5-1% for each 1°C increase in temperature [22], [23]. Crops can be susceptible to damage caused by abrupt decreases in temperature, particularly in the winter and preharvest seasons. An example of frost-induced damage to potato crops in India can reduce crop production by up to 70% at exceptionally low temperatures [24].

Predicting crop yields with statistical models is a standard practice in agricultural research. However, it is tedious and time-consuming [9]. The proliferation of massive data sets (also known as "big data") in recent years has led to the development of cutting-edge analysis methods, the most prominent of which is machine learning. Depending on the specific topic under investigation and the questions posed, one can categorize machine learning models as either descriptive or predictive. Descriptive models are used in data analysis to shed light on the past and provide helpful context. On the contrary, predictive models look ahead to the future [25]. Researchers have used this approach to solve problems in many fields, including medicine, biology, economics, and agriculture [12], [26], [27]. Predicting crop yields with decision support systems is heavily based on machine learning. It is a helpful tool for making informed decisions about what crops to plant and how to care for them during the growing season [28].

This study introduces three machine learning models that aim to estimate the annual yields of six crops (rice, wheat, potatoes, soybeans, sweet potatoes and sorghum) in the Indian region. The statistical analysis carried out on the comprehensive dataset determined that the selected crops exhibit high levels of consumption and cultivation in India. This paper outlines the main contributions in the following manner:

- Collecting crop data for model training and discovering the elements affecting crop productivity.
- Using feature engineering methods to determine which features significantly impact the accurate prediction of crop yield.
- Training of models and examination of hyperparameters to facilitate meaningful interpretation of insights derived from the data.
- Assessment of the model performance using various evaluation metrics.
- Developing the association between crop production and weather parameters.
- A proposal has been made to develop a decision support tool that aims to assist farmers and decision-makers in India in predicting agricultural production. This tool considers various meteorological variables as part of its predictive model. The aim is to improve efforts to tackle climate change and guarantee future food security.
- The efficacy of the suggested model in forecasting agricultural yields is being assessed with other modern techniques employed in crop yield prediction.

This study used three different approaches, each of which utilized a different machine learning algorithm. In order

to make a forecast about the harvest, several techniques, such as multivariate logistic regression, gradient boosting, and k-nearest neighbor, are applied. The coefficient of determination ($R^2$) for these methods is 96.78%, 99.99%, and 98.59%, respectively. Our findings appear among the earliest contributions in this area, especially compared to similar studies conducted in India. Finding reliable climate and agricultural data for India has been challenging for our investigation. Despite this setback, we are confident that the data we acquired would allow us to create a prediction model that neither suffers from overfitting nor underfitting.

We have organized the subsequent portions of this work in the following way. Section II offers a comprehensive review of pertinent scholarly works within agriculture, specifically emphasizing the use of machine learning methodologies to forecast crop yield. Section III comprehensively describes the materials and methods used in the present study. The section consists of an introduction to the topic, a thorough presentation about the proposed crop yield forecast technique, used data source description and the analytical results obtained from it, the implementation of preprocessing and feature engineering methods on the data, an outline of the machine learning models used, a discussion of the metrics used to evaluate the models, and an investigation into the potential utility of cross-validation with hyperparameter optimization. The experimental data are presented in Section IV of this work, followed by a detailed analysis and interpretation of the results. Additionally, the discussion explores the connection between the model's performance and previous publications. The investigation ends in Section VI, which offers several perspectives.

## II. RELATED WORKS

Machine learning (ML), a branch of artificial intelligence that focuses on improving knowledge and abilities through learning algorithms, could improve the accuracy of future crop yield estimates [29]. However, this potential depends on the nature and quality of the data used. Machine learning uncovers hidden relationships and patterns in data. The availability of large data sets has dramatically increased the usefulness of machine learning [30]. The term ''big data'' is used to describe a large amount of data that has been collected quickly from a wide variety of sources. The foundation of machine learning is the development of mathematical models to improve data analysis [31]. Some examples of machine learning techniques used to predict agricultural yields are Ridge Regression, Regression Tree, Support Vector Machine, XGBoost, Convolutional Neural Network, Random Forests, and K-Nearest Neighbor [32], [33]. Ordinary least squares, Least Absolute Shrinkage and Selection Operator, Back Propagation Neural Network, Gaussian process regression, Ensemble Classifiers, Support Vector Machines regression, and Random Forests were among the classification models used and evaluated by Saeed et al. [34].

Numerous studies offer various methods to address the challenges of fluctuating weather conditions in crops. The author of this research [35] critically examines various approaches for evaluating the influence of climate change on agricultural yield and suggests potential adaptation measures through crop models. The article encompasses various farm products, geographical areas, and situations while delving into the difficulties and constraints associated with modeling methodologies. In a separate investigation, a recent scholarly article [36] introduces an innovative system that uses machine learning methodologies to manage meteorological data. The primary objective of this initiative is to improve the overall quality, precision, and practicality of meteorological data, specifically for agricultural purposes. These applications include, but are not limited to, determining crop water requirements, scheduling irrigation, managing pest and disease outbreaks, and predicting crop yields. In a separate investigation [37], a comprehensible machine learning method examines extensive data on climate variables, soil characteristics, and crop yield within the contiguous United States. The text elucidates the various elements that influence crop production and offers valuable perspectives to make informed decisions about sustainable and enduring soil, water, and crop management methods.

Several studies have determined an optimum range for temperature, precipitation, and pesticide needs in which plant growth is maximized using deep learning techniques. A study [38] uses the AquaCrop model to forecast maize production in the eastern province of Rwanda. This model is well known for its focus on water-driven crop simulation. The performance of the model is evaluated in detail using various calibration and validation strategies, using historical weather, pesticides, and crop yield data. The results show that the model effectively captures fluctuations in maize yield, with an Root Mean Square Error (RMSE) of 0.67 t/ha. To forecast maize production in Kenya, the authors of Article [39] use satellite imagery acquired by the Sentinel-2 and Landsat 8 missions in conjunction with machine learning techniques such as Support Vector Machines, Random Forests, and Artificial Neural Networks. The research finds that Support Vector Machines combined with Sentinel-2 data is the best combination, producing a fantastic coefficient of determination ($R^2$) of 0.84 after thorough model comparisons and satellite data evaluations. Article [40] proposes a novel approach to predicting wheat yield using multi-temporal satellite pictures based on deep learning. This study considerably improves wheat yield prediction techniques using a Convolutional Neural Network to extract features from Sentinel-2 images and a recurrent neural network to capture temporal dynamics in wheat growth. The article also compares models extensively and looks at how different input variables affect the precision with which we can predict wheat yield. Although research shows the ability of deep learning and Sentinel-2 satellites to improve crop output prediction, constraints still hinder its wide use. Deep learning algorithms have a high demand for data, necessitating large and costly labeled datasets for practical training. Moreover, the opaqueness of their

"black box" nature impedes interpretability, hence restricting their practical utility for agricultural decision-making. The multispectral resolution of Sentinel-2 data is valuable, but has limitations. The geographical resolution of the system may need to accurately depict the details of small farms or variations within a field, and its reliance on accurate weather data and susceptibility to cloud cover can lead to considerable uncertainty.

Machine learning has emerged as a robust methodology to forecast agricultural productivity in response to dynamic climate patterns. The researchers have integrated various machine learning methods with crop modeling and analyzed complex meteorological factors to offer precise predictions. The intricate interaction among climate change, agricultural practices, and machine learning methodologies presents opportunities for novel approaches to address food security and agricultural sustainability issues [41], [42]. South Asia and India, in particular, need more research on agricultural forecasts. Most previous research relies on forecasts generated using conventional statistical models. Besides, these researches have yet to show when each weather factor is more or less crucial for a given crop. Accurate prediction of future agricultural yields requires access to information from reliable sources. Agricultural data sets are rare owing to the time and effort required to gather and execute the Extract, Transform, and Load (ETL) procedures needed to prepare the data for analysis. Our study aims to create a novel crop yield forecasting model based on machine learning that integrates multiple reliable data sources into their predictions, yielding results highly consistent with empirical evidence.

## III. MATERIALS AND METHODS

To ensure precision, we specify that our study paradigm is quantitative. The emphasis on methodical examination of numerical data is in complete harmony with our objective of constructing a resilient forecast model for crop yields through machine learning. Using a data-driven methodology, we can effectively analyze intricate connections and patterns in extensive datasets, such as meteorological records, pesticide applications, and historical yields. Ultimately, our decision supports progress in data-driven and technology-focused sustainable agriculture.

### A. STUDY AREA

India has approximately 1.27 billion inhabitants, making it the second most populated nation in the world. India, with a total land area of 3.28 million square kilometers, is the seventh largest country in the world. More than 7,500 kilometers of coastline can be found in this particular area. The Himalayas, Thar Desert, Gangetic Delta, and Deccan Plateau contribute to the remarkable agroecological diversity of the country. Western India is home to the Thar desert, whereas the Gangetic delta defines eastern India. Finally, the southern part is where we will find the Deccan Plateau. These aspects of India's topography are mainly responsible for the rich agroecological variety of the country [43], [44]. In this study, the research area encompasses the whole of India.

### B. KEY HARVESTED CROPS

Approximately 49% of the Indian population relies on agriculture as a means of income. The net sown area encompasses 141 million hectares, whereas the gross cropped area spans 195 million hectares. The diverse topography of the country, climatic variations and soil characteristics contribute to the cultivation of a wide range of crops. India cultivates a wide range of tropical, subtropical, and temperate crops [45].

Farmers mainly grow rice as a Kharif crop. It covers approximately one-third of the total cultivated land in India. It caters to the nutritional needs of almost 50% of the Indian population. Rice cultivation is widespread in numerous states in India, with notable concentrations observed in West Bengal, Punjab, and Uttar Pradesh. Other states where rice cultivation is prominent include Tamil Nadu, Assam, and Andhra Pradesh, among others. Rice cultivation requires an annual precipitation range of approximately 150-300 cm, along with deep clay and loamy soil. The mean temperature necessary for the lifetime varies between 21 and 37 °C [46].

Rice is the most important crop in India, with wheat second as the second most prominent agricultural commodity. This particular crop belongs to the Rabi season and is a fundamental dietary component in North and Northeast India. The crop in question is classified as winter and requires low temperatures for optimal growth. Specifically, the ideal temperature range for cultivation spans between 10-15°C during the sowing period and 21-26°C during the harvest period. Wheat exhibits favorable growth patterns within a precipitation range of less than 100 cm and greater than 75 cm. The optimal soil type for wheat farming is characterized by good drainage and high fertility, typically found in loamy and clay soil compositions. Uttar Pradesh, Punjab, and Haryana are the three states with the highest wheat production [47].

### C. DATA COLLECTION

The dataset used in this study for the prediction of crop yield was sourced from the Food and Agriculture Organization (FAO) and the World Data Bank, both publicly accessible [48], [49], [50]. These databases furnish information on annual precipitation, mean annual temperature, pesticide application across various crops, and the outcomes of these applications. The amalgamation of these four datasets, spanning 1990 to 2013, creates a consolidated resource. Our selection of this time frame is based on the availability of authentic and reliable data from reputable sources during the specified period.

### D. PROPOSED YIELD PREDICTION SYSTEM

We introduced a framework rooted in machine learning models. Fig. 1 shows the sequential process of the proposed

crop yield prediction system. The procedure comprises five sequential stages: the Extract, Transform, and Load phase, the preprocessing and feature engineering phase, the hyperparameter tuning and cross-validation phase, the model training phase, the assessment phase, and the concluding phase involving model deployment.

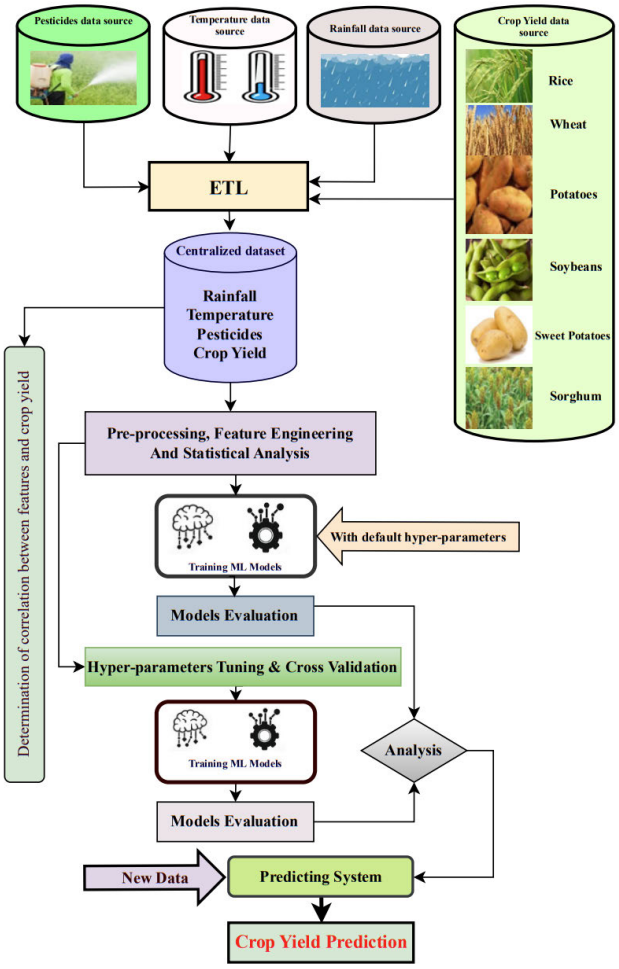



**FIGURE 1.** An overview of the crop yield forecast system.

The initial step of the ETL process involves collecting crop data from various sources. Subsequently, the data sources undergo transformations, cleaning procedures, and processing techniques. Ultimately, the final dataset is consolidated and loaded into a centralized storage system. During the feature engineering phase, we employ various data analysis approaches to examine and comprehend the latent information within the final dataset. The main goal of feature engineering was to preprocess the agricultural dataset to enhance its suitability for integration into machine learning algorithms. During training the models, we employed three different machine learning approaches on the pooled dataset. As a result, a reliable model was created to estimate future harvests.

In particular, multivariate logistic regression, k-nearest neighbor, and gradient boosting models serve as major sources of inspiration for these methodologies. The predictive crop model displays the ability to create exact forecasts for novel crop yields while minimizing errors by employing machine learning generalization. Within the context of supervised learning framework, $h(X, \Theta)$, the machine learning model can be seen as a complicated function. The function uses two arguments: a matrix representing crop data ($X$) and optimization parameters ($\Theta$). Then, it produces an output $Y$. The mathematical representation of this process is provided by(1). We can conceptualize the crop dataset as an assemblage of $(X, Y)$ pairs, where $X$ means a matrix containing $n$ rows and $m$ columns. Each element $x_i$ within it has a numerical value, while $Y \in R^m$ stands as a vector of real numbers. In the proposed model, the matrix $X$ represents the input data, where the columns correspond to the distinct attributes of the crops and the rows correspond to the compilation of time series data. The symbol Y denotes the predicted agricultural output. The variable $m$ denotes the total number of occurrences in the dataset, while the variable $n$ represents the number of characteristics examined. In the present scenario, the number of features is determined to be $n = 10$ due to the feature encoding process.

$$X = [x_1, x_2, \ldots, x_n] \Longrightarrow h(X, \Theta) \Longrightarrow Y \tag{1}$$

The parameters defining the characteristics of our prediction models are widely recognized and readily accessible. These parameters have a significant impact on various agricultural practices within the region. The following items are:

- Temperature (°C): Observations indicate potential yield reductions ranging from 2.5% to 10% across agronomic species during the twentieth century due to increasing temperatures.
- Rainfall, measured in millimeters (mm), refers to the mean annual precipitation in a given geographical region. Water plays a crucial role as a fundamental input in agricultural production, making it susceptible to changes in water availability that can affect agricultural productivity and revenue.
- The pesticide application rate refers to the amount of pesticide, measured in metric tons, used per hectare within a given year. The agricultural sector uses pesticides extensively, particularly in developing nations like India. The primary objective of the use of pesticides is to improve agricultural productivity.
- Yield (hg/ha): the annual yield quantity produced per hectare.

### E. PREPROCESSING AND FEATURE ENGINEERING

The preprocessing and feature engineering stages are essential in the data preparation process for machine learning endeavors. Our collected data set was subjected to a sequence of feature engineering and preprocessing operations prior to training our proposed models.

Upon amalgamating datasets, we identified eight columns, namely 'Unnamed: 0', 'Area', 'Item', 'Year', 'hg/ha_yield',

'average_rain_fall_mm_per_year', 'pesticides_tonnes', and 'avg_temp'. These columns contained 28242 rows of data. The 'Unnamed: 0' column was eliminated, deemed irrelevant to our current research. As our study domain is India, we isolated 4048 rows exclusively about India, leading to the exclusion of the 'Area' column. The 'Item' column comprises categorical information on six crops: potato, rice, sorghum, soybean, sweet potato, and wheat. Eventually, we focused on six columns, encompassing six distinct crops, the years spanning from 1990 to 2013, crop yield (Hg/Ha), average rainfall (mm), and average temperature (°C) for evaluation, thus yielding 4048 rows of information. An essential step in data preprocessing and analysis involves identifying and addressing null values. Null values, indicative of missing data, can undermine data analysis and modeling endeavors' accuracy, reliability, and validity. Vigilant null value handling aids in extracting meaningful insights, making informed decisions, and constructing resilient predictive models. After a comprehensive null value examination, we ascertained the absence of null values in our dataset, allowing us to proceed with this data devoid of nulls.

Detecting and eradicating duplicate records is a pivotal preprocessing step in data analysis and machine learning. Duplicate records involve instances where certain or all attributes of two or more rows in a dataset are identical. This process is indispensable for ensuring data precision, ensuring data quality, and enhancing the reliability of analyses and models. Eliminating duplicate records promotes an accurate depiction of underlying phenomena, supports sound decision-making, and augments the credibility of data-driven undertakings. In our research, we pinpointed 664 instances of duplicated information and subsequently purged them from our dataset. This resulted in 3384 non-duplicated rows across six columns.

During our data exploration, we detected outliers, data points notably distinct from the norm, in the yield and pesticide parameters. Including outliers in data-driven models poses risks, as a single misleading value can alter model-derived conclusions. Detecting and deciding whether to eliminate outliers is crucial. While some extreme values may not necessarily be outliers, others could be the result of errors or artifacts. We identified and removed 242 outlier records using the Inter Quantile Range (IQR) method shown in (2), ultimately yielding 3142 cleaned rows. We also omitted the 'Year' column from the model to exclude its influence during training.

$$\text{IQR} = Q_2 = Q_3 - Q_1 \tag{2}$$

where, $Q_1$ is the cut in the first half of the rank-ordered data set, $Q_2$ is the median value of the set and $Q_3$ is the cut in the second half of the rank-ordered data set.

The ColumnTransformer is a versatile preprocessing tool in machine learning pipelines, streamlining transformations for distinct subsets of columns in datasets. It particularly shines when dealing with mixed numerical and categorical features. Our resultant dataset comprises five columns, with 'Yield' as the target and the remaining four as features. The numerical features ('Avg_rainfall', 'Avg_temp', 'Pesticides') were standardized using StandardScaler, and One Hot Encoding was applied to the categorical feature 'Item', ensuring the appropriate transformation of each subset. Column transformation is calculated using (3)–(6), where (3) is used to perform one hot encoding and (4)–(6) is used to calculate standardization.

$$1_{OHA}(x) := \begin{cases} 1 & \text{if } x \in OHA \\ 0 & \text{if } x \notin OHA \end{cases} \tag{3}$$

Here, OHA is a set of elements and if x is an element of OHA then it returns 1 otherwise returns 0.

$$\text{Mean, } \mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{4}$$

where $y_i$ denotes crop yields.

$$\text{Standard Deviation, } \sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mu_y\right)^2} \tag{5}$$

$$\text{StandardScalar, } z_y = \frac{y - \mu_y}{\sigma_y} \tag{6}$$

Feature engineering, a pivotal and imaginative phase in machine learning, involves refining, transforming, and generating new features from raw data to enhance model performance. It strives to extract pertinent information, capture underlying patterns, and increase feature predictive efficacy. Skillful feature engineering can improve model accuracy, robustness, and generalization.

Table 1 provides a statistical summary of the variables affecting India's crop yields, focusing on temperature and precipitation, the reliance on climatic conditions, and chemicals associated with agricultural actions. Indian temperatures range between 23.26°C and 28.85°C, averaging 26.04°C, with minimal temperature disparities due to favorable climate. Rainfall varies between 935.90 mm and 1401.40 mm, showcasing significant fluctuations due to climatic variability. Rainfall variance is consequential, given that 70% of India's annual rainfall transpires during the monsoon, which is crucial for agriculture. Pesticide use varies widely, from 15075.33 tonnes to 75000.00 tonnes, with notable standard deviation. This variance mirrors the diverse needs for pesticides for different crops.

Multivariate analysis was conducted using Pearson's correlation coefficients (PCE), which revealed the interplay between parameters. Equation (7) is used to calculate the correlation coefficient, $r$. Our findings suggested limited interdependence among these characteristics. Yields exhibited stronger correlations with rainfall and pesticides compared to temperature parameters. This underscores the minimal influence of temperature on yields in the Indian

**TABLE 1.** Statistical analysis of the resultant dataset.

| Criteria | Temperature (°C) | Rainfall (mm) | Pesticides (tonnes) | Yield (hg/ha) |
|---|---|---|---|---|
| **Count** | 3142 | 3142 | 3142 | 3142 |
| **Mean** | 26.04 | 1149.49 | 49116.38 | 62646.28 |
| **Std** | 0.87 | 100.19 | 15075.33 | 71554.31 |
| **Min** | 23.26 | 935.9 | 14485.33 | 6553 |
| **25%** | 25.49 | 1103.3 | 40093.69 | 11998 |
| **50%** | 25.98 | 1165.9 | 46195 | 26972 |
| **75%** | 26.65 | 1207.92 | 61257 | 88368 |
| **Max** | 28.85 | 1401.4 | 75000 | 252815 |

context.

$$\text{PCE}, r = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{7}$$

In this context, $x_i$ represents a specific value about one of the dataset's variables, such as Average Rainfall, while $\bar{x}$ signifies the corresponding mean value for that variable. Similarly, $y_i$ denotes a particular value associated with another dataset variable, like Yield, and $\bar{y}$ represents the mean value for that specific variable. Fig. 2 shows the Pearson's correlation between the variables within our derived dataset.

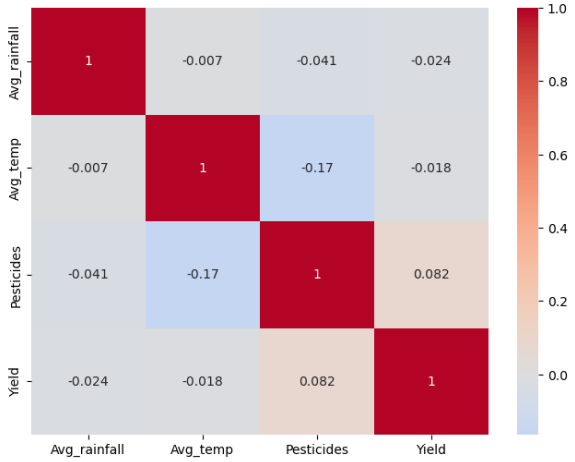



**FIGURE 2.** Pearson correlation between variables.

### F. MACHINE LEARNING MODELS

Numerous scholarly inquiries have amassed substantiating evidence to uphold the contention that Machine Learning is pivotal as an instrumental decision-support mechanism in forecasting agricultural yield. Machine Learning, an advanced technological facet, holds the potential to aid farmers in curtailing agricultural losses by furnishing comprehensive crop guidance and invaluable insights. This ongoing study delves into a plethora of machine learning models, specifically the optimized Yield K-Nearest Neighbours (YK-NN) regressor, Yield Gradient Boosting Regressor (YGBR), and Yield Multivariate Logistic Regression (YMLR). The rationale behind selecting these methodologies stems from the quantitative attributes inherent to the forecasting task, as opposed to qualitative aspects, alongside the scale of the dataset under scrutiny.

#### 1) YIELD K-NEAREST-NEIGHBOR REGRESSOR

The YK-NN Regressor is the initial model devised to anticipate crop yield. This model is rooted in the k-nearest-neighbor methodology and finds its place within supervised learning algorithms. The underlying principle of the YK-NN algorithm rests on the postulation that proximate crop data points share membership within the same category. Its classification as non-parametric underscores its independence from presumptions about the underlying data distribution.

The YK-NN algorithm is renowned for its facile implementation and adeptness in accommodating training data containing intrinsic noise. In this context, the algorithm delineates the procedural framework harnessed by the YK-NN approach to prognosticate crop production values. Our analytical inputs encompass the dataset of crop training instances along with the designated count of neighbors. The YK-NN procedure determines the nearest neighbor for each datum by distance computations involving the focal datum and the remaining data points and iteratively computing the Euclidean distance between the datum $d_i$ and a reference point $d_j$. Ultimately, datum $d_i$ aligns with the cluster of $k$ neighbors, with the majority manifesting akin attributes. The YK-NN algorithm excels when intricate, non-linear interrelations prevail amidst the features and the target variable. This trait results from the algorithm's principal reliance on the resemblance among data points in the feature space.

The YK-NN Regressor algorithm (Algorithm 1) is a versatile and comprehensible method that can predict crop yield, particularly in scenarios including non-linear associations and localized effects. The simplicity of the tool and its capacity to manage missing data make it a valuable instrument for comprehending the variables that influence crop output on a finer-grained and more accurate scale. However, the selection of the parameter k is of significant importance in the context of the YK-NN regression. A minimal value of k can result in overfitting. In contrast, a tremendous value of k can lead to excessive smoothing of the predictions. In this

**Algorithm 1** YK-NN Algorithm

**procedure** SPLIT($_{DF}$,$_{DF}$) ▷ Train $_{DF}$ and Test $_{DF}$

1: **for** each $d_i = x_i, y_i \in \text{Train}_{DF}$ **do**
2: $ED \leftarrow []$ ▷ Initialization for storing Euclidian distance
3: **for** each$_j = (x_j, y_j) \in \text{Train}_{DF} \setminus \{d_i\}$ **do**
4: $ED \leftarrow \sqrt{(d_i - d_j)^2}$
5: **end for**
6: Crop neighbor $d_i \leftarrow$ sorted($ED$)[: $k$]
7: Assign $d_i$ to its nearest crop neighbors' group
8: **end forend procedure**

study, we have used cross-validation and hyperparameter techniques to determine the optimal value of k that leads to the most favorable model performance to predict crop production.

### 2) YIELD GRADIENT BOOSTING REGRESSOR

The YGBR algorithm constructs a collection of weak learners, often decision trees, in which each tree is taught to rectify the mistakes made by the previous tree. The boosting technique employed in this approach involves the successive addition of trees, with each subsequent tree aiming to improve predictions for data points inadequately predicted by the preceding trees. The approach starts with training a rudimentary model, such as a solitary decision tree, using the given dataset. Subsequently, the model performance is assessed and data points exhibiting significant mistakes are identified. Afterward, a subsequent model, in the form of a tree structure, is developed to rectify the inaccuracies produced by the preceding model. As mentioned above, the procedure persists until a predetermined number of trees are incorporated or until a specific threshold of efficacy is attained. The final forecast is obtained by aggregating the predictions of all individual trees. The predictive value of each tree is assigned a weight that is determined by its performance during the training phase.

In addition, the methodology incorporates regularization techniques as a means of minimizing the problem of overfitting. Regularization is a technique that mitigates the complexity of the ensemble by imposing a penalty on significant coefficients within the model. This penalty enhances the model's ability to generalize well to data to which it has not previously been exposed. The algorithm for the YGBR method, which aims to forecast crop production based on the given features of rainfall, temperature, and pesticides used, is presented below.

1) Data Collection and Preparation
   - 1.1 Let DF be the dataset containing crop historical data.
   - 1.2 DF consists of features X and corresponding crop yields
2) Column Transformation, feature scaling, and encoding
   - 2.1 Apply z-score normalization on numerical features such as rainfall, temperature, and pesticides, and transform the columns.
   - 2.2 Apply One hot encoding in the categorical column (items -6 key crops) and transform the resulting columns
3) Data Splitting
   - 3.1 We split DF into a training set $\text{DF}_{\text{train}}$ (80%) and a testing set DF test (20%). As our dataset DF is a moderately sized dataset, so, the $80 - 20$ ratio works well and is a good starting point.
4) Initial Model Training
   - 4.1 4.1 Train the chosen model using the default hyperparameter values in the training data.
   - 4.2 4.2 Train the initial model by calling the subroutine (see Algorithm 2)
   - 4.3 4.3 Evaluate the initial model's performance on a validation set or using cross-validation.
5) Hyperparameter Tuning
   - 5.1 5.1 Find the optimal set of hyperparameters for the model's performance using GridSearchCV.
6) Final Model Training
   - 6.1 6.1 Call the subroutine to retrain the model using the whole training dataset and the optimal hyperparameters.
7) Cross-Validation
   - 7.1 7.1 The training set is cross-validated after model training. This requires splitting the training set into more "folds."
   - 7.2 7.2 The model is validated using each fold and trained using the rest.
   - 7.3 7.3 The performance metrics (such as r2, root mean squared error, and mean absolute error, etc.) are recorded for each fold.
8) Prediction

Within the subroutine, the algorithm begins with a rudimentary forecast of the target $F_0$. Here, Loss $(y_i, \gamma)$ denotes a loss function, and $\min(\gamma)$ represents the sought value for minimizing $\sum_{i=1}^{n}$ Loss $(y_i, \gamma)$. Remarkably, the optimal $\gamma$ that minimizes $\Sigma$ Loss is the mean of y($\bar{y}$) deduced by evaluating the derivative of $\sum_{i=1}^{n}$ Loss $(y_i, \gamma)$ concerning $\gamma$. Step 2.1 involves the calculation of residuals $r_{ik}$, obtained by deriving the loss function for the preceding prediction $F_{k-1}(x)$ and negating the result. In steps 2.2 and 2.3, a regression tree is trained using the feature $x$ against $r$, creating terminal nodes $R_{jk}$. Here, $j$ signifies a terminal node, while $k$ pertains to the tree index.

Step 2.4 entails the pursuit of $\gamma_{jk}$ that minimizes the loss function within each terminal node $j$. The aggregation of loss across all samples $x_i$ belonging to the terminal node $R_{jk}$ is captured by $\sum_{x_i \in R_{jk}}$ Loss $(y_i, F_{k-1}(x_i) + \gamma)$. Thus, $\gamma_{jk}$ represents the ordinary predictive values of regression trees,

characterized as the average of target values (in this context, residuals) within each terminal node.

The combined model's prediction $F_k(x)$ is updated in the ultimate phase. Given a specific x's placement within terminal node $F_k(x) \cdot \gamma_{jk}\left(x \in R_{jk}\right)$ is selected, and the corresponding $\gamma_{jk}$ augments the previous prediction $F_{k-1}(x)$, culminating in the revised prediction $F_k(x)$.

**Algorithm 2** Subroutine

**Require** Initialize the ensemble model $F_o(x)$ with a constant value (e.g., mean of $Y_{\text{train}}$ ) ▷

$F_0(x) = \min(\gamma) \sum_{i=1}^{n} \text{Loss}\,(y_i, \gamma) = \bar{y}$

1: **for** each $k = 1$ *to* $T$ **do**
2: $r_{ik} = -\left[\frac{\partial \text{Loss}\left(y_i, F_{(x_i)}\right)}{\partial F(x_i)}\right]_{F(x)=F_{k-1}(x)}$
3: **for** $i = 1, 2, \ldots, n$ **do**
4: Train regression tree with feature $x$ against $r$,
5: Create terminal node reasons $R_{jk}$
6: **for** $j = 1, 2, \ldots., J_k$ **do**
7: Compute $\gamma_{jk} = \min(\gamma) \sum_{x_i \in R_{jk}} \text{Loss}\,(y_i, F_{k-1}\,(x_i) + \gamma)$
8: **for** $j = 1, 2, \ldots, J_k$ **do**
9: Update the model: $F_k(x) = F_{k-1}(x) + \alpha \sum_{j=1}^{J_k} \gamma_{jk}\left(x \in R_{jk}\right)$
10: **end for**
11: **end for**
12: **end for**
13: **end for**

The crop production forecast frequently entails intricate and non-linear associations among meteorological variables, pesticide use, and crop productivity. The YGBR algorithm demonstrates high effectiveness in capturing non-linear interactions, enabling it to successfully model the intricate complexity inherent in the data. In addition, the method offers a metric for determining the significance of features, thereby highlighting the respective contributions of each component towards the prediction. The analysis of the importance of the characteristics identifies the elements that have the most significant impact on crop production, thus providing valuable insights for decision-making in agricultural practices. Additionally, it mitigates the potential for overfitting and enhances the model's capacity for generalization. This practice seems particularly advantageous in situations with a scarcity of crop production data, as it helps mitigate the risk of overfitting. In general, the YGBR algorithm demonstrates versatility and robustness as a regression algorithm, making it suitable for the prediction of crop yields. This algorithm effectively utilizes ensemble learning techniques and proficiently manages intricate associations among various features.

#### 3) YIELD MULTIVARIATE LOGISTIC REGRESSION

The suggested YMLR model can perform both regression and classification tasks as a supervised machine learning approach. In contrast to multivariate linear regression, the proposed crop model deals with situations in which the outcome variable (dependent variable) takes on a dichotomous form. Using a logistic function, YMLR links independent and dependent variables. Multivariate logistic regression is represented mathematically by (8).

$$Y_{\text{yield}} = \frac{1}{1 + e^{-[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + - - - \bar{+} \beta_n x_n]}} \tag{8}$$

Here, the logistic function is applied to the linear combination $\theta 0 + \theta 1 \times 1 + \theta 2 \times 2 + \cdots + \theta nxn$ for computing the crop yield target's response. In this Equation, $\Theta = [\theta 0, \theta 1, \ldots \theta n]$ denotes the vector of unknown accurate parameters, while $x1, x\ 2, \ldots xn$ (e.g., Avg_rainfall, Avg_temp, pesticides, Item) are independent variables. $Y$ signifies the target, explanatory variable, or dependent variables.

When the data is split into a training and testing set, the crop model is trained on all training data instances xi. Learning parameters are optimized using Gradient Descent because the method tends to converge to a global minimum. In order to assess the efficiency of the crop yielding model and the model's performance, we compute the loss at each time step using (9).

$$\log_{\text{loss}} = -\frac{1}{K} \sum_{i=1}^{K} y_i \cdot \log\left(\hat{y}_i\right) + (1 - y_i) \cdot \log\left(1 - \hat{y}_i\right) \tag{9}$$

where $\hat{y} = h\,(\beta_i, x_i)$.

### G. K-FOLD CROSS-VALIDATION WITH HYPERPARAMETERS TUNING

Parameters are the model's internal variables refined using information gleaned via training. They determine how the model represents the data's structure. For example, the coefficients of the predictor variables are the parameters in a linear regression model. However, hyperparameters are parameters whose values are determined beforehand, rather than during, the learning process. These cannot be learned from the data but are set at the outset. They aid in regulating the learning procedure and have a sizeable impact on the model's efficiency. The maximum depth of a decision tree is an example of a hyperparameter.

Compared to a simple train-test split, the regression model's performance estimate provided by K-fold cross-validation is more accurate. The bias-variance tradeoff in the model can then be more easily determined. Poor training and validation set performance of a model with solid bias may necessitate a more sophisticated model or feature engineering. Overfitting may occur if a model with a significant variance performs well on the training set but poorly on the validation set. With a more complete picture of the model's performance over various subsets of the data,

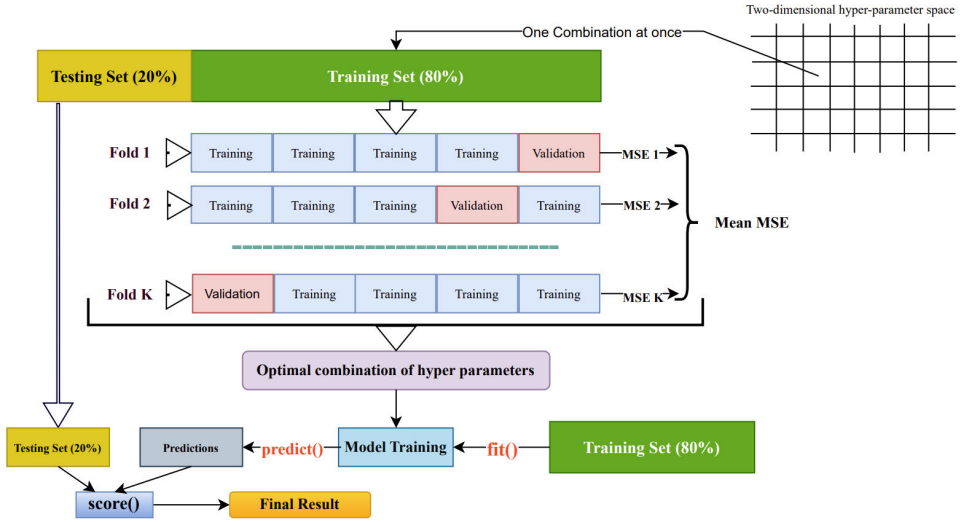



**FIGURE 3.** Hyper-parameter tuning with GridsearchCV.

**TABLE 2.** Details of the implementation platform.

| Name | Details | |
|---|---|---|
| Processor | Intel(R) Core(TM) i7-4500U CPU @ 1.80GHz, 2401 MHz, 2 Core(s), 4 Logical Processor(s) | |
| Make | ACER Aspire E1-572 | |
| Architecture | 64 | |
| Operating System | Windows 10 | |
| Memory Allotted | 4 GB | |
| GPU | Iris Xe | |
| Coding Language | Python | |
| Framework | Anaconda/Jupiter Notebook v6.5.1 | |
| Python | numpy, pandas, mathplotlib, sesborn | |
| Scikit-learn | sklearn.preprocessing | OneHotEncoder, StandardScaler |
| | sklearn.compose | ColumnTransformer |
| | sklearn.model_selection | train_test_split, cross_val_score, KFold, GridSearchCV |
| | sklearn.linear_model | LinearRegression |
| | sklearn.neighbors | KNeighborsRegressor |
| | sklearn.ensemble | GradientBoostingRegressor |
| | sklearn.metrics | mean_absolute_error, mean_squared_error, r2_score |
| Sci-py | scipy.stats | pearsonr |

k-fold cross-validation aids in finding a middle ground between these two concerns. A model overfitting or underfitting the data can also be detected. It is a sign of overfitting if the model performs well in the training but poorly in the validation set. Underfitting is likely if performance in both training and validation is poor. Adjusting the complexity and regularization of a model is facilitated by understanding the trade-off between training and validation performance.

In machine learning, GridSearchCV is a method to find the optimal hyperparameters for a given algorithm through a systematic search. It aids in pinpointing the sweet spot of a model's hyperparameters for maximum efficiency. GridSearchCV is effective because it uses cross-validation to assess model performance after exploring every possible combination of hyperparameters. To avoid overfitting, models can be tested on different subsets of the training data using cross-validation.

Hyper-parameter tuning with GridsearchCV is shown in Fig. 3. Before beginning model training, the dataset was split in half (i.e., into train and test sets), with the former containing 80% of the data and the latter 20% serving as validation. We attempted several combinations, beginning with a split of 50% training/test, increasing the training component, decreasing the test portion, and so on. Finally, we discovered that a ratio of 0.80:0.20 between training and testing yielded the best outcomes. The Python GridSearchCV

program was used to fine-tune more parameters, and its output was used to train the models and provide accurate predictions.

### H. IMPLEMENTATION PLATFORM

The present investigation into the forecasting of crop yields is executed using a standalone computer system, employing the Jupyter Notebook framework. Python is employed to script the three suggested machine learning models. A comprehensive account of the configuration of the implementation framework can be found in Table 2.

## IV. RESULTS

The present study utilizes three optimized models, namely YMLR, YK-NN, and YGBR, which leverage historical data from the United Nations Food and Agricultural Organization and the World Bank datasets. These models evaluate the agricultural output of crucial crops in the context of India. After the model's training with historical data, an evaluation is performed using a collection of data instances that have not been encountered before, often referred to as test samples. Loss, which denotes the disparity between anticipated and actual values within the test samples, is used to gauge the model's efficiency.

### A. MODEL EVALUATION

Mean Absolute Error (MAE), RMSE, and R-squared ($R^2$) were used as evaluation metrics for the machine learning models' performance. MAE, which measures the most significant deviation from the target value, is the most common performance measure. A low MAE value for the model indicates excellent performance.

The RMSE is a regularly utilized metric for analyzing the effectiveness of prediction models, notably in regression assignments. Mean differences, or residuals, between expected and observed values are calculated. Due to its clarity, interpretability, and ability to capture the severity of forecast mistakes, RMSE gives a thorough study of regression model performance, pinpointing areas for improvement and streamlining the model selection process.

The coefficient of determination, also known as $R^2$, is a final metric that measures the prediction accuracy of a regression model. In summary, the coefficient of determination ($R^2$ score) evaluates the degree of concordance between the forecasts (about crop data) and the fundamental hypothesis of the predictive models. The $R^2$ takes values between 0 and 1, representing a score. A score of 1 signifies a high level of accuracy in crop prediction, indicating that the model performs exceptionally well. On the other hand, a score of 0 shows a static model that lacks predictive capability and guesses the average of the answers in the training set. Equation (10) calculates the mean absolute error, which is used to assess the performance of the model. Equation (11), on the other hand, provides a mechanism to determine the root mean square error, which also aids in evaluating model performance. Lastly, Equation (12) facilitates the calculation of the coefficient of determination, commonly referred to as r-squared, as another pivotal metric for gauging the effectiveness of the models.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \left| y_{\text{pred}} - y_{\text{act}} \right| \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left( y_{\text{pred}} - y_{act} \right)^2} \tag{11}$$

$$R^2 = 1 - \frac{\sum \left( y_{\text{pred}} - y_{\text{act}} \right)^2}{\sum \left( y_{\text{pred}} - y_{\text{avg}} \right)^2} \tag{12}$$

where $y_{\text{act}}$ is the actual crop yields; $y_{\text{pred}}$ denotes the predicted crop yields and $y_{\text{avg}}$ represents the mean of crop yields.

Table 3 displays the assessment measures for the machine learning models used in their default configuration. The evaluation encompasses both the training and testing datasets.

**TABLE 3. ML models evaluation with default hyper-parameters.**

| Models | MAE (t/ha) | | RMSE (t/ha) | | $R^2$(%) | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| YMLR | 0.3202 | 0.331 | 0.542 | 0.5582 | 97.02 | 96.76 |
| YK-NN | 0.0719 | 0.0921 | 0.4518 | 0.4982 | 99.51 | 99.11 |
| YGBR | 0.0827 | 0.0873 | 0.1146 | 0.1183 | 99.94 | 99.13 |

Compared to the YK-NN and YMLR models, the YGBR model performed exceptionally well. MAE was 0.0827t/ha on the training dataset and 0.0873 $t/$ ha on the testing dataset when the YGBR model was trained using the default settings of learning_rate $=$ 0.1, n_estimators $=$ 100, and min_samples_split $=$ 2. The training dataset MAE for the YMLR model was 0.3202t/ha, whereas the testing dataset MAE was 0.3310 $t/$ ha. Meanwhile, with its default settings (n_neighbors $=$ 5 ), the YK-NN model achieved an MAE of 0.0719t/ha on the training dataset and 0.0921t/ha on the testing dataset.

YK-NN outperformed the competition in the training dataset, while there was some variation in performance between the training and testing evaluation metrics within the MAE criterion. However, when looking at the test dataset, YGBR had the smallest MAE value.

Based on the regression analysis conducted on the training dataset, it was found that the YGBR model exhibited the most effective performance, as indicated by its $R^2$ score of 99.94%. On the contrary, the YMLR model exhibited the lowest performance with an $R^2$ score of 97.02%. The YGBR and YK-NN models demonstrated similar predictive performance in the test dataset, evident from their almost identical $R^2$ values of 99.11% and 99.13%, respectively. These observations suggest the presence of overfitting in the model.

To address this concern, we implemented a strategy for tuning hyper-parameters to determine an optimal model configuration that appropriately aligns with the dataset while

mitigating the risk of overfitting. We executed this initiative by applying a cross-validation methodology.

### B. HYPERPARAMETER TUNNING WITH CROSS-VALIDATION

The data in Table 3 suggest the possibility of overfitting. We used the cross-validation method to adjust the hyperparameters and find the best setting for our model, ensuring it was consistent with the data.

The cross-validation approach splits the dataset into K-folds. The technique involves training the model on some partitions while using another partition as the test set. The aforementioned iterative procedure is carried out for each part of the dataset that serves as a test case. Therefore, the effectiveness of the resulting model is evaluated by computing its mean performance.

The GridsearchCV package was used to facilitate this procedure. The machine learning model (the estimator), a hyperparameter grid, and the desired number of subsets (either a k-fold or cross-validation value) are only some of the inputs that can be provided to the library. After that, the library gives us the best estimator and the optimal values for its hyperparameters.

We partitioned the dataset using an 80/20 ratio, where 80% of the data was allocated for training within the cross-validation process, and the remaining 20% was set aside for testing. This deliberate partitioning strategy ensures a robust evaluation of the model's performance while mitigating the potential for overfitting.

**TABLE 4.** K-Folds with GridsearchCV.

| Folds | KNeighborsRegressor | | GradientBoostingRegressor | |
|---|---|---|---|---|
| | RMSE | $R^2$ | RMSE | $R^2$ |
| 2 | 8272.88 | 0.972859 | 376.27 | 0.999943 |
| 3 | 5673 | 0.986991 | 301.69 | 0.999962 |
| 4 | 6202.01 | 0.984327 | 315.22 | 0.999959 |
| 5 | 5421.06 | 0.987922 | 267.04 | 0.999972 |
| 6 | 4441.01 | 0.992035 | 252.87 | 0.999974 |
| 7 | 4217.03 | 0.992528 | 241.37 | 0.999977 |
| 8 | 4798.97 | 0.989892 | 249.37 | 0.999975 |
| 9 | 3752.73 | 0.993948 | 254.07 | 0.999974 |
| 10 | 3766.04 | 0.993441 | 234.12 | 0.999978 |
| 11 | 3715.57 | 0.993642 | 223.79 | 0.99998 |
| Mean | 5171.63 | 0.988758 | 271.58 | 0.999969 |
| Std. deviation | 1432.71 | 0.006456 | 46.59 | 1.14E-05 |

Table 4 summarizes the results of an examination of two distinct regression models: the KNeighborsRegressor (with n_neighbors set to 3 ) and the GradientBoostingRegressor (with learning_rate = 0.5, n_estimators = 150 and random_state = 42 ). Multiple cross-validation folds are used to calculate the RMSE and the R square $R^2$ values. Each cell in the table represents one cross-validation fold. The RMSE and $R^2$ values of both models across all folds are displayed in the table. The RMSE quantifies how off the model is on average, whereas the correlation coefficient $R^2$ evaluates how well the model fits the data.

The 'Mean' row provides the average values of RMSE and $R^2$ across all folds, indicating the overall performance of the models. The 'Standard deviation' row provides information about the variability or spread of the RMSE and $R^2$ values across different folds, which can help assess the consistency of the model's performance.

The results show that the GradientBoostingRegressor consistently outperforms the KNeighborsRegressor in terms of both RMSE and $R^2$ across all folds. The values in the 'Mean' row indicate that the GradientBoostingRegressor has a lower average RMSE and a higher average $R^2$, suggesting better predictive accuracy and fit to the data than the KNeighborsRegressor.

Additionally, the minor standard deviations for the GradientBoostingRegressor's RMSE and $R^2$ values indicate that its performance is relatively consistent across different folds, implying good generalization ability. Overall, the hyperparameter with GridsearchCV provides insights into the comparative performance of the two regression models based on the provided evaluation metrics.

Results from adjusting model hyperparameters and running sensitivity analyses are summarized in Table 5. The parameters were iteratively optimized to arrive at the values shown in the table. Regarding accuracy, the YGBR model performed 99.99% on the training data with an MAE of 0.0160t/ha and 99.98% on the test data with an MAE of 0.0182 t/ha. Similarly, the YK-NN model exhibited a training data score of 99.36% along with an MAE of 0.0380t/ha and a test data score of 98.29% with an MAE of 0.0769t/ha.

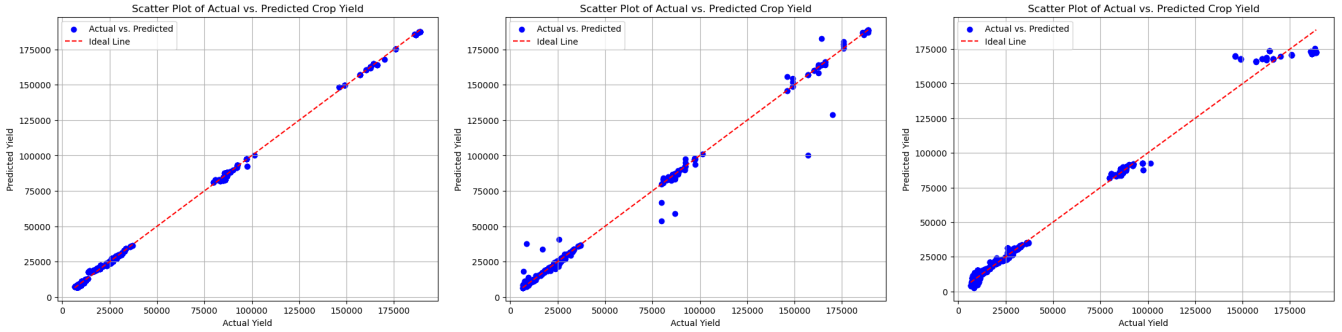
The data presented in Table 5 shows that the standard deviations of the training and testing results are minimal after the hyper-parameter tuning achieved via GridSearchCV. This phenomenon signifies the enhanced generalizability of the trained models.

**TABLE 5.** Results obtained from the models after parameter adjustment.

| Models | MAE (t/ha) | | RMSE (t/ha) | | $R^2$(%) | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| YMLR | 0.2624 | 0.2719 | 0.4551 | 0.4569 | 96.78 | 96.13 |
| YK-NN | 0.038 | 0.0769 | 0.3486 | 0.4612 | 99.36 | 98.29 |
| YGBR | 0.016 | 0.0182 | 0.0233 | 0.0255 | 99.99 | 99.98 |

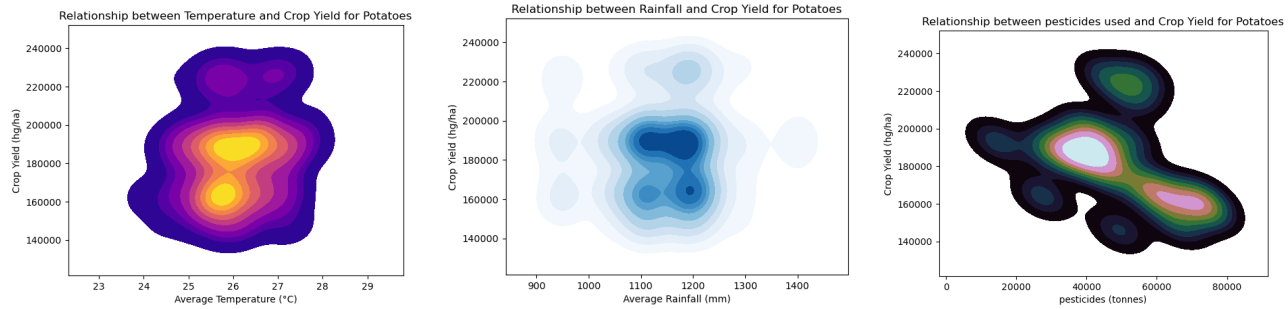
### C. CORRELATION BETWEEN PREDICTED AND ACTUAL RESPONSES

The correlation between predicted and actual crop yields is shown in the three graphs presented in Fig. 4. The positive linear regression trend is shown in the scatter plot. The existence of specific data points that display minor deviations from the central cluster is noteworthy, as these variations might be linked to intrinsic biases within the models. However, the scatter plot's depiction of the data indicates the presence of a linear correlation between the parameters, indicating the possibility of decreasing predictive variability.




(a) YGBR (b) YK-NN (c) YMLR

**FIGURE 4.** The relationship between the expected and actual production of crops: (a) YGBR (b) YK-NN (c) YMLR.




(a) Correlation of temperature and crop yield (b) Correlation of rainfall and crop yield (c) Correlation of pesticides and crop yield

**FIGURE 5.** Correlation between independent variables and yield for the potato crop.

A noteworthy observation from Fig. 4 is that the optimized YGBR model appears to outperform the YK-NN and YMLR models in predictive accuracy and alignment with actual crop yields.

### D. CORRELATION BETWEEN CROP YIELDS AND CLIMATE CONDITIONS

The investigation of the intricate connections between meteorological factors, notably rainfall, temperature, the use of pesticides in the cultivation of significant crops, and the resulting yields, has been meticulously carried out. This exploration has been facilitated by applying the Kernel Density Estimate (KDE) tool, an analytical instrument adept at uncovering the inherent correlations between diverse variables. The diligent deployment of this tool has allowed us to discern the intricate interrelationships that prevail among rainfall, temperature, pesticide application, and yields attained for each of the critical crops under study.

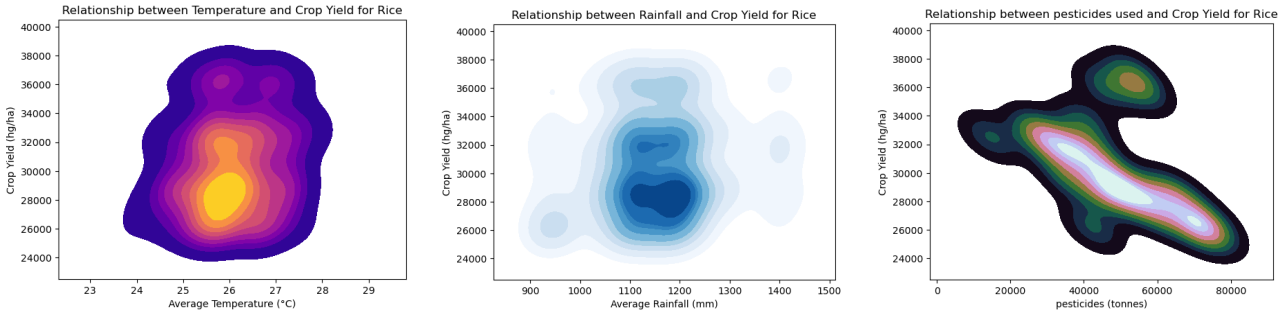
We meticulously constructed Figs. 5–10 to depict and communicate these intricate correlations. These figures substantiate the correlations between the specified meteorological parameters and the resulting crop yields for the prominent crops of interest. Through a thoughtful synthesis of data visualization and analytical techniques, the presented figures effectively elucidate the profound associations that weather variables and pesticide use hold with the agricultural productivity of key crops.

In Fig. 5, a comprehensive analysis reveals that attaining the highest potato yield (1, 90, 000hg/ha) is associated with specific optimal conditions. These conditions include an average temperature of 26.2°C, an average rainfall of 1120 mm, and the utilization of approximately 3300 tons of pesticides. The discernible relationship between these parameters and potato yield underscores the importance of precise environmental management strategies.

Fig. 6 provides a different perspective, showcasing the primary factors in achieving the optimal rice yield (32,000 hg/ha). The intricate interplay of variables becomes apparent since the most favorable conditions entail an average temperature of 25.8°C, an average rainfall of 1200mm, and the application of 3900 tonnes of pesticides. The intricate alignment of these parameters accentuates their crucial roles in improving rice productivity.

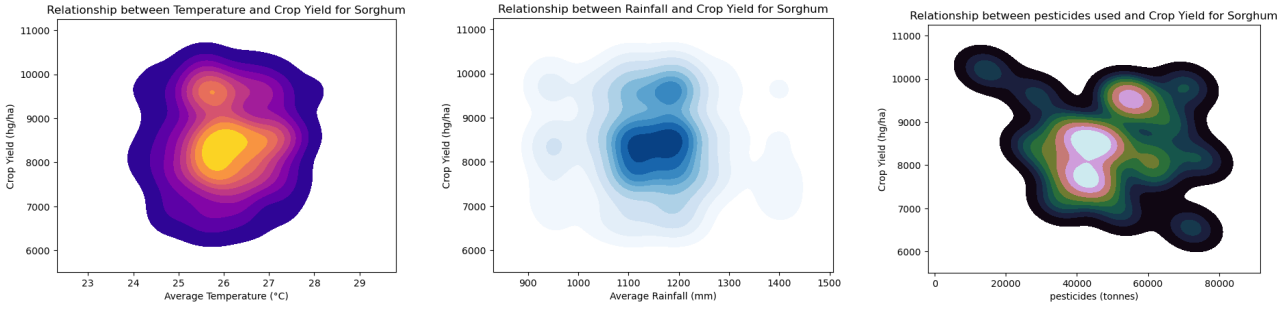
Moving to Fig. 7, an insightful portrayal emerges concerning the factors underpinning optimal sorghum yield (9,800 hg/ha). The intricate fabric of yield enhancement is woven with precision, entailing an average temperature of 25.5°C, an average rainfall of 1100mm, and a notable usage of 57000 tonnes of pesticides. This intricate synergy under-scores the nuanced relationship between environmental conditions and sorghum productivity.

Fig. 8 provides a distinct lens through which the pursuit of the highest soybean yield (11,200 hg/ha) comes into focus. The intricate orchestration of elements becomes apparent,



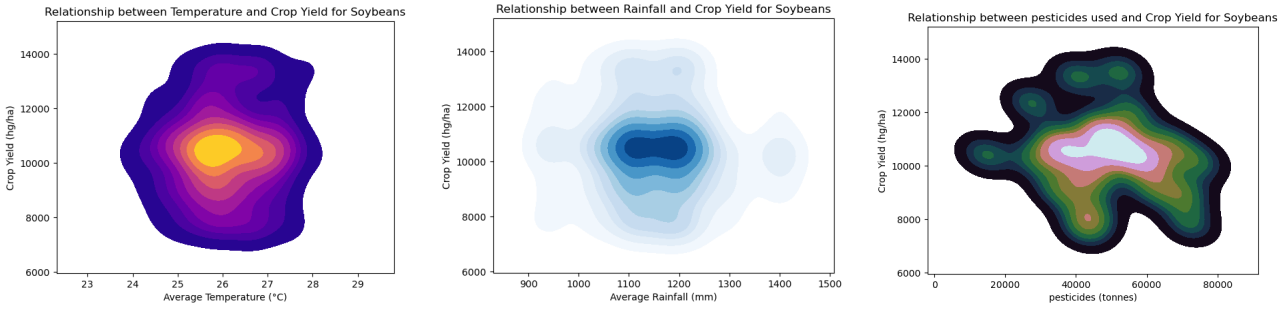

(a) Correlation of temperature and crop yield (b) Correlation of rainfall and crop yield (c) Correlation of pesticides and crop yield

**FIGURE 6.** Correlation between independent variables and yield for the rice crop.




(a) Correlation of temperature and crop yield (b) Correlation of rainfall and crop yield (c) Correlation of pesticides and crop yield

**FIGURE 7.** Correlation between independent variables and yield of the sorghum crop.




(a) Correlation of temperature and crop yield (b) Correlation of rainfall and crop yield (c) Correlation of pesticides and crop yield

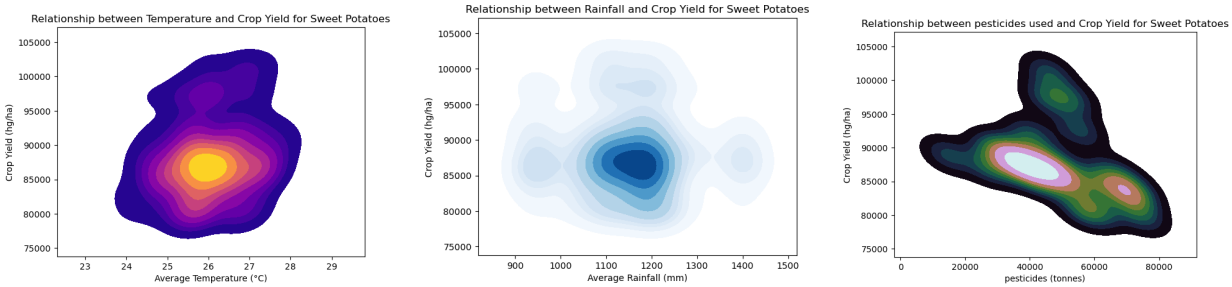
**FIGURE 8.** Correlation between independent variables and yield for the soybean crop.

as optimal parameters encompass an average temperature of 25.6 °C, an average rainfall of 1270mm, and an application of 48000 tonnes of pesticides. The discernible patterns in these parameters illuminate the way to maximize soybean productivity.

In Fig. 9, a discernible narrative unfolds, detailing the essential parameters for achieving the highest yield of sweet potatoes (90, 000hg/ha). The nuanced interplay of conditions takes center stage, with the optimal configuration comprising an average temperature of 26°C, an average rainfall of 1100 mm, and the careful application of 35000 tonnes of pesticides. This intricate interplay underscores the importance of tailored environmental adjustments.

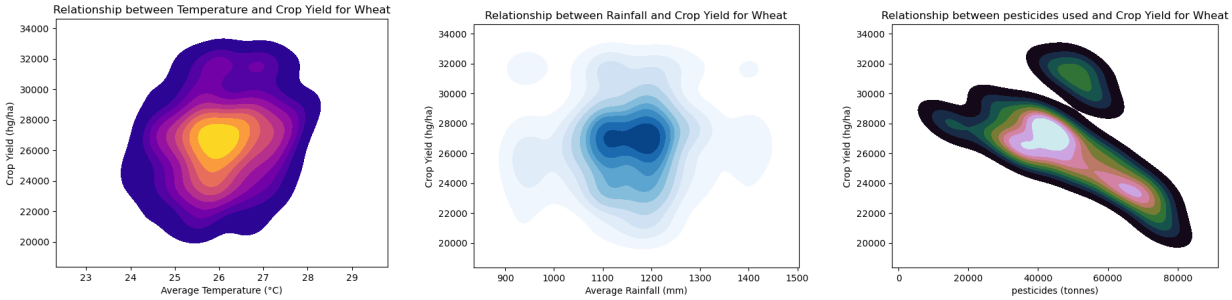
Fig. 10 expands the narrative by presenting a distinctive perspective on the pursuit of optimal wheat yield (28, 500hg/ha). The intricate balance of factors becomes apparent, as attaining the highest yield involves an average temperature of 25.5°C, an average rainfall of 1, 180 mm, and the use of 40,000 tonnes of pesticides. This interwoven relationship underscores the strategic amalgamation of environmental variables to maximize wheat productivity.

In summary, each figure uniquely depicts the intricate interplay of environmental parameters and their critical role in achieving the highest crop yields. These insights highlight the importance of tailored strategies to optimize agricultural productivity while considering the nuanced relationships




(a) Correlation of temperature and crop yield (b) Correlation of rainfall and crop yield (c) Correlation of pesticides and crop yield

**FIGURE 9.** Correlation between independent variables and yield for the sweet potato crop.




(a) Correlation of temperature and crop yield (b) Correlation of rainfall and crop yield (c) Correlation of pesticides and crop yield

**FIGURE 10.** Correlation between independent variables and yield for wheat crop.

between temperature, rainfall, pesticide application, and crop yield.

### E. COMPARATIVE ANALYSIS OF THE PROPOSED MODELS

Fig. 11 compares the $R^2$ scores of three different machine learning models (YGBR, YK-NN, and YMLR). A regression model's $R^2$ score, often called its coefficient of determination, reflects the amount of variation in the dependent variable that can be accounted for by the independent variables. Values closer to 1 indicate a more accurate representation of the data by the model.

On the training data, the YGBR model scored an $R^2$ of 99.99, while on the testing data, it scored 99.98. With such a high $R^2$ value, the YGBR model is a good match for the training and testing datasets, allowing for reliable predictions of the dependent variable.

When applied to the training data, the YK-NN model achieved an $R^2$ of 98.59; on the testing data, it achieved an $R^2$ of 97.29. Although lower than YGBR, these $R^2$ values show that the YK-NN model matches both datasets well.

The $R^2$ score for the YMLR model was 96.78 in the training data and 96.13 in the testing data. Based on these results, the YMLR model matches the datasets reasonably well but has a poorer fit than YGBR and YK-NN.

In conclusion, Fig. 11 emphasizes the differences in $R^2$ scores across the three machine learning models on the training and testing datasets. With the most significant $R^2$ values, YGBR demonstrates superior prediction abilities and competence. Although their $R^2$ ratings are lower, $YK-NN$ and YMLR also show respectable performance.

## V. DISCUSSION

A combination of factors, controllable and uncontrollable, influences crop yields. The initial considerations encompass various criteria, such as the selection of crop or seed types, the practice of tillage, the application of fertilizers, and numerous other variables. Uncontrollable factors encompass environmental variables that are outside of human control. These factors encompass but are not restricted to, elements such as precipitation, ambient and soil temperatures, atmospheric humidity, soil moisture, solar radiation, and comparable parameters. Should any of these attributes exceed or fall below the plant's optimal range, it can hinder its development, consequently affecting its productivity. Using the aforementioned meteorological variables is essential for an accurate crop production forecast. However, the accessibility and availability of the relevant data may limit the extent to which this can be achieved. Therefore, this study used rainfall, temperature, and pesticide data to forecast the yield of six prominent crops, including rice, potatoes, wheat, soybeans, sweet potatoes, and sorghum, during the year in India.

Investigating crop production prediction using machine learning techniques is paramount because it provides valuable insights into productivity patterns, facilitating informed decision-making processes. This study involved the utilization of a dataset containing information on rainfall, pesticide usage, temperature, and crop yield. The dataset was employed to train three models to identify the most effective yield predictor. The findings of this study will be valuable for recommending a yield predictor to system developers and informing future research endeavors.

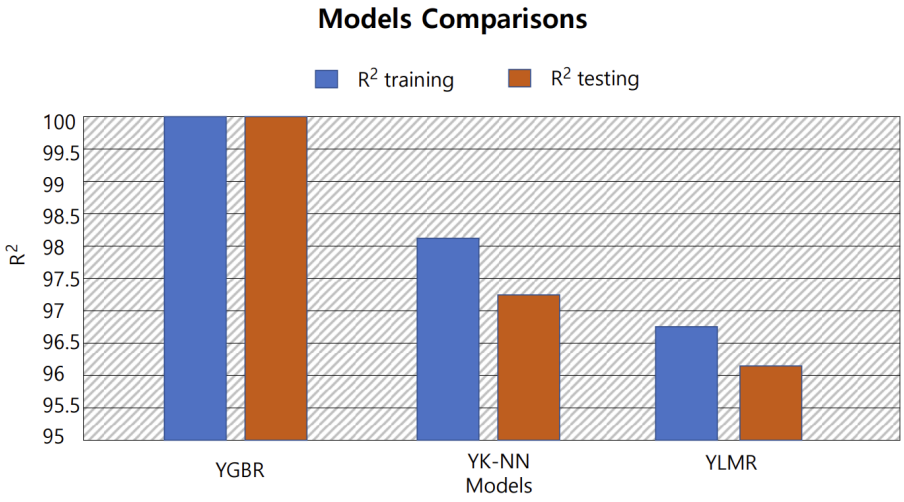



**FIGURE 11.** Comparison of proposed models.

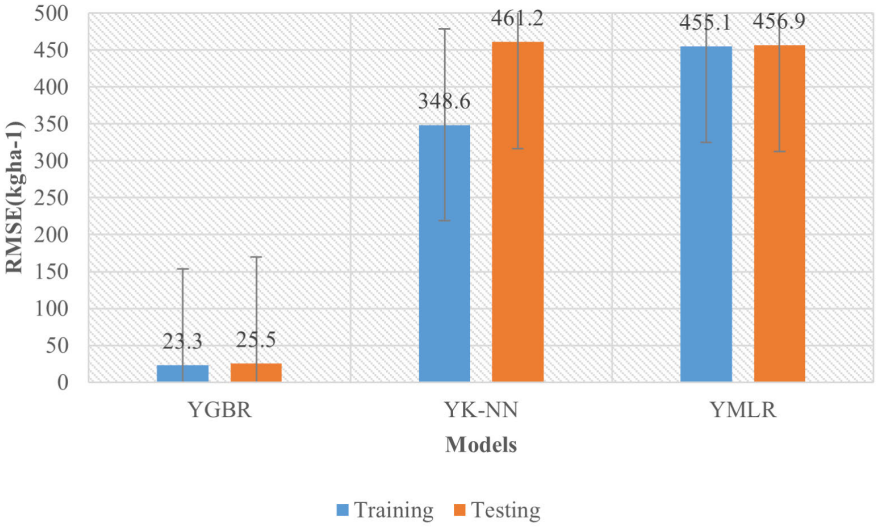



**FIGURE 12.** Models' comparison using RMSE.

Researchers have used various climate-related indicators, particularly rainfall and temperature, to predict agricultural production. This propensity arises due to the limited availability of alternative variables, such as soil moisture, solar radiation, humidity, and other factors. The findings derived from the evaluated models demonstrate that the precision values for forecasting the yield, measured by the RMSE, were 25.5, 461.2, and 456.9 kgha-1 for the YGBR, YK-NN, and YMLR models, respectively. The findings suggest that the YGBR model performs better, as illustrated in Fig. 12.

The performance results suggest that the most suitable model for constructing an early crop yield prediction system is YGBR, which utilizes an optimized gradient boosting technique. Furthermore, the gradient boosting model is a prominent machine learning approach in contemporary agricultural production forecasting. Previous research has demonstrated favorable performance outcomes through the use of several models, such as support vector regression, Artificial Neural Networks, Convolutional Neural Network, deep neural networks, and long- and short-term memory. However, researchers consciously make decisions regarding the choice of machine learning models and predictors (features), considering diverse aspects such as the dataset's attributes, the nature of the dependent variable (target yield), the dataset's scale, and the accessibility of the data. In this study, the climatic elements were considered, recognizing the significant impact of climate change on agricultural output. Crop growth can be impacted by climatic change at different stages, potentially leading to a decrease in yield. The use of rainfall, pesticides, and temperature was based on the careful consideration of the growth of the crops, serving as the underlying rationale for our approach. Despite yielding satisfactory predictive outcomes, the YGBR model's performance was hindered in this study due to the lack of data

concerning additional parameters like soil moisture, solar radiation, and humidity. The inclusion of these characteristics would likely lead to improved predictive outcomes. The upcoming study will examine these aspects employing the Internet of Things (IoT) to gather meteorological and hydraulic variables. Subsequently, we will use these variables as input for the YGBR model to predict crop output.

## VI. CONCLUSION

The present study introduces a framework to predict agricultural yield, using an optimal gradient boosting machine learning model. The framework has been developed explicitly for the cultivation of key crops, such as rice, wheat, potatoes, soybeans, sweet potatoes, and sorghum, in the specific context of India. The successful execution of the study was facilitated by integrating data obtained from reputable sources, including the FAO of the United Nations and the Climate Knowledge Portal of the World Bank. Data acquisition involved collecting information about agriculture, pesticides, and weather from various databases. These diverse data sources were integrated by implementing Extract, Transform, and Load (ETL) procedures, resulting in increased efficiency and a more simplified workflow. After the completion of the integration process, a series of preprocessing approaches, analytical techniques, and feature engineering procedures were applied to enhance the amalgamated dataset. The analysis aimed to gain a deeper understanding of the latent information included within the data. The approach used in this study required a thorough data analysis and investigation of the interrelationships between the variables. We used transformation and encoding techniques on the variables. Normalization was performed on the data to achieve standardization of their scale. The manipulation of the variables aided in creating the data for the model's training. The present study aimed to assess the predictive capacities of three machine learning models, namely the YMLR, YK-NN, and YGBR models, in the context of forecasting agricultural output. The preliminary training results of these models demonstrated acceptable performance, although they still needed to meet the anticipated standard. Consequently, the models were improved by using hyperparameter tuning and cross-validation methodologies. The results of the model assessment showed that, out of the three models, the YGBR model had the highest score. The test data demonstrate a coefficient of determination ($R^2$) value of 99.98% and a MAE of 0.0182 t/ha. When we compare the YMLR and YK-NN models, we can observe that they exhibit $R^2$ values of 96.13% and 97.29%, respectively. The model has shown diminished predictive inaccuracy, as evidenced by the RMSE value of 0.0233 t/ha.

Furthermore, the training period of the model can be deemed suitable. The suggested prediction models demonstrate a high degree of generalizability and possess the capacity to handle large-scale datasets effectively. This study contributes significantly to the continuing digital revolution in agriculture by optimizing agricultural methods and improving production while limiting resource use. Machine learning models are practical foundational elements of pervasive computing systems because they extract substantial insights from the data gathered. The findings of this study indicate that the Gradient Boosting model had superior performance compared to the Linear regression and K-Nearest Neighbor models for predicting the yields of principal crops. Our next study aims to examine the development of a framework architecture that combines the proposed YGBR model with the IoT technologies to forecast yields. This project will incorporate the IoT technology to monitor precipitation, air temperature, and some extreme climate events in real-time.

## REFERENCES

[1] M. Kavita and P. Mathur, "Crop yield estimation in India using machine learning," in *Proc. IEEE 5th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Oct. 2020, pp. 220–224.

[2] M. A. A. Osman, J. O. Onono, L. A. Olaka, M. M. Elhag, and E. M. Abdel-Rahman, "Climate variability and change affect crops yield under rainfed conditions: A case study in Gedaref state, Sudan," *Agronomy*, vol. 11, no. 9, p. 1680, Aug. 2021.

[3] M. N. Thimmegowda, M. H. Manjunatha, L. Huggi, H. S. Shivaramu, D. V. Soumya, L. Nagesha, and H. S. Padmashri, "Weather-based statistical and neural network tools for forecasting rice yields in major growing districts of Karnataka," *Agronomy*, vol. 13, no. 3, p. 704, Feb. 2023.

[4] C. Song, W. Ma, J. Li, B. Qi, and B. Liu, "Development trends in precision agriculture and its management in China based on data visualization," *Agronomy*, vol. 12, no. 11, p. 2905, Nov. 2022.

[5] M. Chandler. (2023). *How Does Climate Change Affect Agriculture?* Accessed: May 12, 2023. [Online]. Available: https://impakter.com/how-climate-change-affects-agriculture/

[6] L. Burrows. (Sep. 2022). *A Better Understanding Of Crop Yields Under Climate Change*. Accessed: May 12, 2023. [Online]. Available: https://seas.harvard.edu/news/2022/09/better-understanding-crop-yields-under-climate-change,

[7] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan, "Crop yield prediction using machine learning and deep learning techniques," *Proc. Comput. Sci.*, vol. 218, pp. 406–417, Jan. 2023.

[8] D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, "Machine learning for large-scale crop yield forecasting," *Agricult. Syst.*, vol. 187, Feb. 2021, Art. no. 103016.

[9] R. Affoh, H. Zheng, X. Zhang, W. Yu, and C. Qu, "Influences of meteorological factors on maize and sorghum yield in Togo, West Africa," *Land*, vol. 12, no. 1, p. 123, Dec. 2022.

[10] S. Javadinejad, S. Eslamian, and K. O. A. Askari, "The analysis of the most important climatic parameters affecting performance of crop variability in a changing climate," *Int. J. Hydrol. Sci. Technol.*, vol. 11, no. 1, pp. 1–25, 2021.

[11] L. Liu and B. Basso, "Impacts of climate variability and adaptation strategies on crop yields and soil organic carbon in the US midwest," *PLoS ONE*, vol. 15, no. 1, Jan. 2020, Art. no. e0225433.

[12] A. Wegrzyn, A. Klimek-Kopyra, E. Dacewicz, B. Skowera, W. Grygierzec, B. Kulig, and E. Flis-Olszewska, "Effect of selected meteorological factors on the growth rate and seed yield of winter wheat—A case study," *Agronomy*, vol. 12, no. 12, p. 2924, Nov. 2022.

[13] J. Cao, Z. Zhang, F. Tao, L. Zhang, Y. Luo, J. Zhang, J. Han, and J. Xie, "Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches," *Agricult. Forest Meteorol.*, vol. 297, Feb. 2021, Art. no. 108275.

[14] Q.-U.-A. Ahmad, H. Biemans, E. Moors, N. Shaheen, and I. Masih, "The impacts of climate variability on crop yields and irrigation water demand in South Asia," *Water*, vol. 13, no. 1, p. 50, Dec. 2020.

[15] V. Geethalakshmi, R. Gowtham, R. Gopinath, S. Priyanka, M. Rajavel, K. Senthilraja, M. Dhasarathan, R. Rengalakshmi, and K. Bhuvaneswari, "Potential impacts of future climate changes on crop productivity of cereals and legumes in Tamil Nadu, India: A mid-century time slice approach," *Adv. Meteorol.*, vol. 2023, pp. 1–17, Jan. 2023.

[16] D. Chakraborty, S. Saha, B. K. Sethy, H. D. Singh, N. Singh, R. Sharma, A. N. Chanu, I. Walling, P. R. Anal, S. Chowdhury, S. Hazarika, V. K. Mishra, P. K. Jha, and P. V. V. Prasad, "Usability of the weather forecast for tackling climatic variability and its effect on maize crop yield in northeastern Hill region of India," *Agronomy*, vol. 12, no. 10, p. 2529, Oct. 2022.

[17] P. S. Nishant, P. Sai Venkat, B. L. Avinash, and B. Jabber, "Crop yield prediction based on Indian agriculture using machine learning," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Jun. 2020, pp. 1–4.

[18] M. D. M. Kadiyala, S. Nedumaran, J. Padmanabhan, M. K. Gumma, S. Gummadi, S. R. Srigiri, R. Robertson, and A. Whitbread, "Modeling the potential impacts of climate change and adaptation strategies on groundnut production in India," *Sci. Total Environ.*, vol. 776, Jul. 2021, Art. no. 145996.

[19] P. Aggarwal, S. Vyas, P. Thornton, B. M. Campbell, and M. Kropff, "Importance of considering technology growth in impact assessments of climate change on agriculture," *Global Food Secur.*, vol. 23, pp. 41–48, Dec. 2019.

[20] P. Ekanayake, W. Rankothge, R. Weliwatta, and J. W. Jayasinghe, "Machine learning modelling of the relationship between weather and paddy yield in sri Lanka," *J. Math.*, vol. 2021, pp. 1–14, May 2021.

[21] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," in *Proc. 13th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, 2016, pp. 1–5.

[22] D. J. Reddy and M. R. Kumar, "Crop yield prediction using machine learning algorithm," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, 2021, pp. 1466–1470.

[23] Y. J. N. Kumar, V. Spandana, V. Vaishnavi, K. Neha, and V. Devi, "Supervised machine learning approach for crop yield prediction in agriculture sector," in *Proc. 5th Int. Conf. Commun. Electron. Syst. (ICCES)*, 2020, pp. 736–741.

[24] M. R. Yadav, M. Choudhary, J. Singh, M. K. Lal, P. K. Jha, P. Udawat, N. K. Gupta, V. D. Rajput, N. K. Garg, C. Maheshwari, M. Hasan, S. Gupta, T. K. Jatwa, R. Kumar, A. K. Yadav, and P. V. V. Prasad, "Impacts, tolerance, adaptation, and mitigation of heat stress on wheat under changing climates," *Int. J. Mol. Sci.*, vol. 23, no. 5, p. 2838, Mar. 2022.

[25] A. Simon, P. I. Moraru, A. Ceclan, F. Russu, F. Chetan, M. Bărdas, A. Popa, T. Rusu, A. I. Pop, and I. Bogdan, "The impact of climatic factors on the development stages of maize crop in the Transylvanian plain," *Agronomy*, vol. 13, no. 6, p. 1612, Jun. 2023.

[26] X. Han, L. Chang, N. Wang, W. Kong, and C. Wang, "Effects of meteorological factors on apple yield based on multilinear regression analysis: A case study of Yantai Area, China," *Atmosphere*, vol. 14, no. 1, p. 183, Jan. 2023.

[27] L. Zhang, Z. Zhang, F. Tao, Y. Luo, J. Cao, Z. Li, R. Xie, and S. Li, "Planning maize hybrids adaptation to future climate change by integrating crop modelling with machine learning," *Environ. Res. Lett.*, vol. 16, no. 12, Dec. 2021, Art. no. 124043.

[28] G. Morales, J. W. Sheppard, P. B. Hegedus, and B. D. Maxwell, "Improved yield prediction of winter wheat using a novel two-dimensional deep regression neural network trained via remote sensing," *Sensors*, vol. 23, no. 1, p. 489, Jan. 2023.

[29] P. Muruganantham, S. Wibowo, S. Grandhi, N. H. Samrat, and N. Islam, "A systematic literature review on crop yield prediction with deep learning and remote sensing," *Remote Sens.*, vol. 14, no. 9, p. 1990, Apr. 2022.

[30] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agricult.*, vol. 177, Oct. 2020, Art. no. 105709.

[31] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction," *IEEE Access*, vol. 9, pp. 63406–63439, 2021.

[32] A. A. Deshmukh, A. Srivatsa, A. Monteiro, and C. Gajakosh, "Crop yield prediction to achieve precision agriculture using machine learning," in *Proc. IEEE 2nd Int. Conf. Mobile Netw. Wireless Commun. (ICMNWC)*, Dec. 2022, pp. 1–6.

[33] S. Agarwal and S. Tarar, "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms," *J. Phys., Conf.*, vol. 1714, no. 1, Jan. 2021, Art. no. 012012.

[34] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Frontiers plant Sci.*, vol. 10, p. 621, Jan. 2019.

[35] D. Xiao and W. Shi, "Modeling the adaptation of agricultural production to climate change," *Agriculture*, vol. 13, no. 2, p. 414, Feb. 2023.

[36] C. E. Hachimi, S. Belaqziz, S. Khabba, B. Sebbar, D. Dhiba, and A. Chehbouni, "Smart weather data management based on artificial intelligence and big data analytics for precision agriculture," *Agriculture*, vol. 13, no. 1, p. 95, Dec. 2022.

[37] D. Sihi, B. Dari, A. P. Kuruvila, G. Jha, and K. Basu, "Explainable machine learning approach quantified the long-term (1981–2015) impact of climate and soil properties on yields of major agricultural crops across CONUS," *Frontiers Sustain. Food Syst.*, vol. 6, Apr. 2022, Art. no. 847892.

[38] C. Rugimbana, "Predicting maize (Zea mays) yields in eastern province of Rwanda using aquacrop model," Ph.D. dissertation, Dept. Meteorol., Univ. Nairobi, Nairobi, Kenya, 2019.

[39] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN framework for crop yield prediction," *Frontiers Plant Sci.*, vol. 10, p. 1750, Jan. 2020.

[40] G. Impollonia, M. Croci, A. Ferrarini, J. Brook, E. Martani, H. Blandinières, A. Marcone, D. Awty-Carroll, C. Ashman, J. Kam, A. Kiesel, L. M. Trindade, M. Boschetti, J. Clifton-Brown, and S. Amaducci, "UAV remote sensing for high-throughput phenotyping and for yield prediction of miscanthus by machine learning techniques," *Remote Sens.*, vol. 14, no. 12, p. 2927, Jun. 2022.

[41] Q. Sun, Y. Zhang, X. Che, S. Chen, Q. Ying, X. Zheng, and A. Feng, "Coupling process-based crop model and extreme climate indicators with machine learning can improve the predictions and reduce uncertainties of global soybean yields," *Agriculture*, vol. 12, no. 11, p. 1791, Oct. 2022.

[42] M. Kuradusenge, E. Hitimana, D. Hanyurwimfura, P. Rukundo, K. Mtonga, A. Mukasine, C. Uwitonze, J. Ngabonziza, and A. Uwamahoro, "Crop yield prediction using machine learning models: Case of Irish potato and maize," *Agriculture*, vol. 13, no. 1, p. 225, Jan. 2023.

[43] *India at a Glance*. Accessed: Oct. 9, 2023. [Online]. Available: https://www.fao.org/india/fao-in-india/india-at-a-glance/en/

[44] India Brand Equity Foundation, *Agriculture in India: Information About Indian Agriculture & Its Importance*, New Delhi, India, Sep. 2023.

[45] R. Bhatt and R. S. Meena, "Delineation of soil moisture potentials and moisture balance components," in *Soil Moisture Importance*. London, U.K.: IntechOpen, 2020.

[46] A. K. Bhardwaj, D. Rajwar, U. K. Mandal, S. Ahamad, B. Kaphaliya, P. S. Minhas, M. Prabhakar, R. Banyal, R. Singh, S. K. Chaudhari, and P. C. Sharma, "Impact of carbon inputs on soil carbon fractionation, sequestration and biological responses under major nutrient management practices for rice-wheat cropping systems," *Sci. Rep.*, vol. 9, no. 1, p. 9114, Jun. 2019.

[47] A. Mahajan and R. D. Gupta, *Integrated Nutrient Management (INM) in a Sustainable Rice-Wheat Cropping System*. Berlin, Germany: Springer, 2009.

[48] W. Bank. *Open Data*. Accessed: Oct. 9, 2023. [Online]. Available: https://data.worldbank.org/

[49] FAO. *Dataset*. Accessed: Oct. 9, 2023. [Online]. Available: https://www.fao.org/home/en/

[50] R. Patel. *Crop Yield Prediction Dataset*. Accessed: Oct. 9, 2023. [Online]. Available: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset

**MD JIABUL HOQUE** received the B.Sc. degree in CSE from Chittagong University of Engineering and Technology (CUET), and the M.Sc. degree in CSE from the University of Greenwich, London, U.K. He is currently a prominent Faculty Member of the Faculty of Science and Engineering, International Islamic University Chittagong, and a distinguished Ph.D. Research Fellow with CUET. With a strong focus on machine learning and the IoT, he has authored over 20 research articles in prestigious international journals and conferences. He actively participates in workshops, seminars, and symposiums in these fields, contributing significantly to academic and scientific progress. His dedication and expertise are driving advancements in these domains, making him a recognized figure in the global research community.

**MD. SAIFUL ISLAM** received the B.Sc. degree in electrical and electronic engineering from Chittagong University of Engineering and Technology (CUET), Chittagong, Bangladesh, in 2010, and the M.S. and Ph.D. degrees from the University of Ulsan, South Korea. He is currently an Associate Professor with the Department of Electronics and Telecommunication Engineering, CUET. His research interests include artificial intelligence, signal processing, image and video processing, fault detection, and diagnosis. The above research lines have produced more than 50 publications on international journals, presentations within international conferences, and book chapters. He also acts as a referee in several highly reputed journals and international conferences.

**JIA UDDIN** received the M.Sc. degree in telecommunications from Blekinge Institute of Technology, Sweden, in 2010, and the Ph.D. degree in computer engineering from the University of Ulsan, South Korea, in January 2015. He is currently an Assistant Professor with the AI and Big Data Department, Endicott College, Woosong University, South Korea. He was an Assistant Professor with the CSE Department, BRAC University, and the CCE Department, International Islamic University Chittagong, Bangladesh. He was a member of the Self-Assessment Team (SAC) of CSE, BRACU, in the HEQEP project funded by the World Bank and the University of Grant Commission Bangladesh, from 2016 to 2017.

**MD. ABDUS SAMAD** (Member, IEEE) received the Ph.D. degree in information and communication engineering from Chosun University, South Korea. He was an Assistant Professor with the Department of Electronics and Telecommunication Engineering, International Islamic University Chittagong, Chattogram, Bangladesh, from 2013 to 2017. He has been a Research Professor with the Department of Information and Communication Engineering, Yeungnam University, South Korea. His research interests include signal processing, antenna design, electromagnetic wave propagation, the applications of artificial neural networks, and millimeter-wave propagation by interference and atmospheric causes for 5G and beyond wireless networks. He won the Prestigious Korean Government Scholarship (GKS) for his doctoral study.

**BEATRIZ SAINZ DE ABAJO** is currently an Associate Professor with the Department of Signal Theory, Communications and Telematics Engineering, University of Valladolid, Spain. Her research interests include the development and evaluation of e-Health systems, m-Health, medicine 2.0., and cloud computing. She belongs to the GTe Research Group and focuses on topics related to electronic services for the information society. Among the lines of research, the group works to develop innovative solutions in the field of health that help patients improve their quality of life and facilitate the work of health professionals.

**DÉBORA LIBERTAD RAMÍREZ VARGAS** is currently with Universidad Europea del Atlántico, Santander, Spain. She is also associated with Universidad Internacional Iberoamericana, Campeche, Mexico, and Universidad de La Romana La Romana, Dominica.

**IMRAN ASHRAF** received the M.S. degree (Hons.) in computer science from Blekinge Institute of Technology, Karlskrona, Sweden, in 2010, and the Ph.D. degree in information and communication engineering from Yeungnam University, Gyeongsan-si, South Korea, in 2018. He was a Postdoctoral Fellow with Yeungnam University, where he is currently an Assistant Professor with the Information and Communication Engineering Department. His research interests include positioning using next-generation networks, communication in 5G and beyond, location-based services in wireless communication, smart sensors (LIDAR) for smart cars, and data analytics.