IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 July 2024, accepted 2 September 2024, date of publication 4 September 2024, date of current version 13 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3454511

RESEARCH ARTICLE

An Ensemble Machine Learning Framework for Cotton Crop Yield Prediction Using Weather Parameters: A Case Study of Pakistan

SYED TAHSEEN HAIDER[1,2], WENPING GE[2], JIANQIANG LI[1], (Senior Member, IEEE), SAIF UR REHMAN[3], AZHAR IMRAN[4], (Senior Member, IEEE), MOHAMED ABDEL FATTAH SHARAF[5], AND SYED MUHAMMAD HAIDER[3]

[1]School of Software Engineering, Beijing University of Technology, Beijing 100124, China
[2]School of Computer Science and Technology, Xinjiang University, Ürümqi, Xinjiang 830046, China
[3]University Institute of Information Technology, Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi 46300, Pakistan
[4]Faculty of Computing and Artificial Intelligence, Air University, Islamabad 44000, Pakistan
[5]Industrial Engineering Department, College of Engineering, King Saud University, Riyadh 11421, Saudi Arabia

Corresponding authors: Saif Ur Rehman (saif@uaar.edu.pk) and Azhar Imran (azhar.imran@au.edu.pk)

This work was supported by King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Program under Grant RSPD2024R704.

ABSTRACT In Pakistan, agriculture is one of the most common and least lucrative professions. It provides between 18% and 25% of Pakistan's overall gross domestic product (GDP). The majority of Pakistan's crops, like cotton, are completely weather-dependent. Regarding this, farmers are constantly attempting to implement new techniques and technology to boost crop yields. Technology-based approaches to crop yield analysis, such as machine learning (ML) and data mining, are causing a boom in the agricultural sector by altering the revenue scenario through the growth of the best crop. By utilizing ML algorithms to analyze agriculture climatic data, it is possible to increase crop yields. The proposed research was carried out in two dimensions. Initially, field observations were made to determine the effects of daily variations in meteorological parameters, such as rainfall, temperature, and wind, on plant growth and development at each phonological stage of cotton crop production. Throughout the Kharif Seasons 2005-2020, various phonological stages of the cotton crop grown in the fields of the Ayyub Agriculture Research Institute in Faisalabad (Central Punjab) were monitored using meteorological and phonological observations, as well as soil data. Finally, the cotton prediction framework as Random Forest Extreme Gradient (RFXG) has been proposed to predict cotton production based on observed data. RFXG concentrates on the quantification of machine learning algorithms and their practical application. The workings of RFXG have been divided into two phases. In the very first phase of data collection, preprocessing, attribute selection, and data splitting have been presented. In the following phase, prediction and evaluation were developed. The comparative results show that the prediction results of the proposed RFXG using the optimization algorithm are significantly improved by 0.05 RMSE (Root Mean Square Error) in comparison to the traditional Extreme Gradient Boost (XGB) model, which has a RMSE of 0.07. Proposed technique also compared with some baseline approaches of cotton predication. Comparison shows that proposed technique achieves better results as compared to baseline approaches. The proposed RFXG model (ensemble-based method) can bag, stack, and boost, making it fast and efficient predications as compared with existing approaches. Bagging averages, the results of numerous decision tree fit to different subsets of the same dataset to increase accuracy. The proposed study will be very useful in the future to close the gap between the current yield obtained and the potential yield of this cultivar, which is grown in Pakistan and other cotton-growing locations.

INDEX TERMS Cotton, machine learning, climatic change, yield, weather.

I. INTRODUCTION

Agriculture is one of the most popular occupations in Pakistan, but also being one of the least paid. It contributes

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

between 18% and 25% of Pakistan's total Gross Domestic Product [1]. The vast bulk of Pakistan's agricultural products, such as cotton, rely entirely on the weather. Farmers are constantly looking for new methods and technology to increase the number of crops they harvest, especially cotton. Technology-based approaches to crop yield analysis, such as machine learning and data mining, are causing a boom in the agricultural sector by modifying the revenue scenario by cultivating the most profitable crop. This led to an increase in the number of people working in the agricultural industry [2]. It is thus conceivable to improve crop yields by using algorithms for machine learning to analyze agricultural climate data.

Cotton, an essential raw material for numerous textile mills and ginning plants nationwide, underpins the livelihoods of many farmers and individuals throughout the cotton value chain. Nevertheless, cotton production encounters hurdles owing to seasonal shifts, extreme weather events such as floods and droughts, and fluctuations in meteorological parameters, including temperature, rainfall, and soil moisture [3]. These variations precipitate a decline in crop yields, introducing substantial challenges to several economic sectors. Notably, the agricultural sector is highly susceptible to these alterations, with atmospheric conditions impacting approximately 60% of its productivity. Researchers are diligently exploring the repercussions of climate change on agriculture and investigating potential adaptive strategies. While acclimating to temperature alterations is perceived as a viable mitigation method, low-income countries—particularly those with a pronounced dependence on agriculture—are anticipated to confront escalated challenges in adapting to these shifts moving forward [4].

Climate change poses the greatest threat to humanity and wildlife, escalating global food insecurity as temperatures rise. Vulnerable crops like cotton and wheat face substantial risks due to these temperature increases, potentially causing a significant reduction in agricultural production, particularly in underdeveloped nations, according to the Intergovernmental Panel on Climate Change (IPCC) [5]. Pakistan's agricultural sector may suffer adverse effects from increased temperatures, precipitation changes, and rising sea levels, exacerbated by global warming and drought [6]. The country is prone to climate-related disasters [7], with biological responses to climate change, particularly higher temperatures, being highly impactful, as highlighted in the IPCC's Special Report on limiting global warming to 1.5 degrees Celsius. Climate-induced shifts in temperature and rainfall patterns could detrimentally affect the growth, development, timing, and yields of various crops, including cotton. Pakistan's cotton production faces the risk of negative impacts, and extreme weather events globally reduce agronomic agriculture production by 50% [5]. Although several climate change studies have focused on Pakistan's wheat, rice, and maize crops, cotton's vulnerability remains less understood, emphasizing the urgency of addressing its potential detrimental effects. Agriculture plays a pivotal role in Pakistan's economy, heavily reliant on rainfall patterns that vary from arid to semi-arid, making farmers highly sensitive to changes in the social and economic climate. Pakistan has experienced intense rainfall, tornadoes, severe droughts, earthquakes, and recurring floods, as reported by the Task Force on Climate Change (TFCC) [8]. Weather parameters such as temperature, wind speed, wind direction, and rainfall significantly impact agriculture and various aspects of life.

The significance of research on cotton production prediction using machine learning techniques cannot be overstated. Farmers, politicians, and other stakeholders engaged in the cotton production process must have an accurate and timely forecast of cotton output. Accurate yield forecasts enable farmers to make educated choices on planting, harvesting, and selling of crops. Precise estimates may also help them manage the risks associated with weather occurrences and other variables that might affect agricultural output. For policymakers, precise projections of cotton output may help them make educated judgements concerning agricultural policies, such as subsidies and price support programs. Precise forecasts also enable them to identify regions of the nation that are likely to face cotton crop shortfalls and to take the necessary steps to alleviate the effects of such shortages. Accurate forecasts of cotton output may assist textile businesses in scheduling their production and ensuring a sufficient supply of raw materials. Consumers may plan their purchases and guarantee they have access to the things they need if cotton crop projections are correct.

In addition to the practical advantages of precise estimates of cotton output, there is tremendous intellectual value in this field of study. Using machine learning methods in agricultural research is an interesting new area in the world of agriculture. The creation and enhancement of machine learning models for cotton yield prediction may aid in the comprehension of the intricate interactions between weather patterns, soil conditions, and other variables that affect crop yields. Overall, research on cotton output prediction using machine learning approaches is of high importance and has the potential to generate large advantages for a broad variety of stakeholders. By enhancing our capacity to properly anticipate cotton yields, we can assist farmers, policymakers, textile producers, and consumers in making better informed choices, mitigating risks associated with weather events, and ensuring a consistent supply of raw materials for the textile sector.

A. RESEARCH CONTRIBUTIONS

The contributions of this study encompass several key challenges associated with the utilization of machine learning methods, such as Random Forest and XGBoost, for predicting cotton yield based on weather parameters. This research contributes by tackling the formidable challenge of obtaining reliable and comprehensive datasets that encompass historical weather data and corresponding cotton yield information for various cities in Pakistan. The dedicated effort and resources invested in data collection and quality assurance enhance the foundation of this predictive model.

Another significant contribution lies in the identification of the most relevant weather parameters that exert a substantial influence on cotton yield. This study offers valuable insights into feature selection, drawing upon domain knowledge and agricultural expertise to pinpoint the appropriate weather variables, thereby ensuring the accuracy of yield predictions. This research acknowledges and addresses the issue of imbalanced datasets, where the distribution of samples among different yield classes (high yield, moderate yield, low yield) significantly varies. By devising strategies to mitigate this imbalance, the study contributes to developing a model that does not exhibit bias towards the majority class, thus enhancing the overall performance and fairness of the predictive model.

This paper will discuss the following section. The literature review is covered in section II, the methodology is covered in section III, the experiment and findings are covered in section IV, and the conclusion and future work are included in section V.

II. LITERATURE REVIEW

This section provides a comprehensive overview of the background and the work done in the proposed research. This work also discussed Cotton Production, the Cotton Growing process in Pakistan, and Strategies for increasing cotton yield; machine learning and predictive analysis on the cotton production process have also been elaborated. The effects of the basics of climate change and weather parameters are also defined briefly.

Crop yields [9] and wheat yields have been modelled with remarkable success, surpassing traditional statistical methods. In a study conducted in Australia, machine-learning approaches were employed to identify increasing episodes of heat as a primary driver of yield losses. Based on these findings, the author concludes that machine learning can facilitate enhanced analysis and the identification of correlations between various predictors and climate conditions, even under challenging circumstances [10]. Moreover, it can generate high-resolution data and enable investigations into the drivers of climate change. Weather conditions are linked to the prevalence of cotton pests and diseases. Cotton pests and diseases are more likely to occur during autumn and winter when temperatures are favorable, humidity is low, rainfall is scarce, wind speed is low, sunlight duration is long, and evaporation is low [11]. Moreover, the factors impacting cotton pests and diseases differ by region. Cotton grown in different locations may be subject to more complex developmental factors.

Pest and disease incidence in cotton is not solely determined by climatic variables but also by other factors such as cotton growth, pest life cycles, and so forth, with climate playing only a partial role. The model developed by Qing Xin Xiao took a different approach by excluding the pest value feature while training various models. In a study conducted by the author [12], it was noted that understanding the patterns of pest and disease occurrence plays a crucial role in model learning. However, Xiao's model focused solely on evaluating meteorological elements and past pest data without considering the law of pest and disease incidence. Surprisingly, this model could still train different models successfully.

On the other hand, in [13] model showed reasonable predictions across various datasets. However, it was observed that incorporating historical data on pests and diseases could significantly enhance the model's performance, particularly for Whitefly pests, as evidenced by its higher AUC (Area under the Curve) score of 0.9257. Therefore, including historical pest and disease data in the model could greatly improve its accuracy and effectiveness in predicting pest incidents. This study aims to identify the most effective model for crop prediction with the ultimate aim of assisting farmers in selecting crops based on meteorological conditions and existing soil nutrients. This study evaluates several well-known algorithms based on Gini and Entropy, two distinct criteria. According to the findings, Random Forest delivers the highest level of accuracy among the three.

In the research [14], a comparison between the DT (decision tree) model and the SVM (Support vector machine) models revealed that the SVM model surrenders the correlation coefficient between the actual and forecast cotton yield of the DT model, as the DT model was much higher. This was discovered due to the comparison between the DT model and the SVM (Support vector machine) models. Furthermore, the RMSE (Root Mean Square Error) value was discovered to be significantly lower than the value obtained from SVM estimation. Cotton yield can be predicted with 84% accuracy using the DT model. In a similar study for wheat production, it was calculated as 91%. Because there are weak relationships between the data, the model created with SVM has a very low accuracy. The author discovered that even with such data, high-accuracy predictions could be made using DT modeling [15]. Using a DT and SVM model, the author believes that more precise results can be acquired by performing a study among Diyarbakir cotton producers and all adjacent producers.

The impact of climate change on crops, characterized by an increase in air temperature and alterations in precipitation patterns, has disrupted Brazil's current crop forecasting methods [16]. However, the low accuracy of these forecasts can be attributed to the approach's failure to account for climate change. Considering this, modeling techniques such as machine learning have been proposed as an alternative to yield forecasting. Agro-meteorological models and machine learning algorithms have proven reliable in reproducing the interactions between climatic components and agricultural yield variability [17]. Machine learning, a technology for data analysis, seeks to automate the development of analytical models. It can potentially make data processing faster, more efficient, and more precise [18]. Machine learning algorithms have become the standard approach for predicting agricultural yields. For instance, studies have utilized random forests for the accurate forecasting of sugarcane yield, RF for the forecasting of maize yield, and the integration of machine learning and climatic data in the forecasting of soybean yield using a satellite in Southern Brazil [19]. However, there is a lack of research on applying machine learning methods in predicting cotton production.

This comprehensive review of agricultural yield prediction models examines the application of machine learning techniques in forecasting crop yields, particularly for cotton and wheat. The study highlights the successful implementation of machine learning in surpassing traditional statistical methods, focusing on identifying climate-related drivers of yield losses. Notably, various models are compared, emphasizing the importance of incorporating historical pest and disease data for enhanced accuracy, and Random Forest emerges as a leading performer in crop prediction. The research also underscores the impact of climate change on crop forecasting. It suggests that machine learning, with its data analysis and automation capacity, is becoming the standard for predicting agricultural yields. The investigation extends its contributions by developing a user-friendly smartphone application, a GPS-based location identifier for rainfall estimates, and a program recommending optimal fertilizer application times. Ultimately, the findings demonstrate that machine learning, particularly the Decision Tree method, can achieve high accuracy in crop yield predictions, offering valuable insights for assisting farmers in crop selection based on meteorological conditions and soil nutrients. The study addresses the challenges posed by the increasing number of parallel flow requests in Data Center Networks (DCNs). The knapsack model treats link bandwidth as the knapsack capacity and incoming flows as items, with the PSO algorithm optimizing which flows to forward first based on their size and value. This approach is designed to improve performance metrics such as flow completion time, packet loss rate, and overall good put, particularly under conditions of high parallel flow detection [28] Auto-encoders are particularly advantageous over other models like Restricted Boltzmann Machines (RBMs) due to their ability to learn more complex data representations. Unlike RBMs, which are generative models designed to learn probability distributions, auto-encoders consist of an encoder and a decoder that facilitate the reconstruction of input data, making them more tractable and easier to train for various applications [29].

The growth of applications and flow demands in Data Center Networks presents challenges for efficient flow management. Each flow is defined by size and value, determined by flow extraction and type of service (ToS) values. Flow sizes are categorized by ToS decimal values, prioritizing higher values. Particle Swarm Optimization (PSO) optimizes flow management by addressing dense parallel flow scenarios, keeping network links below a 70% load. Experimental results demonstrate the method's superiority for varied flow sizes [30]. Autoencoders are a pivotal focus in unsupervised learning, prized for their capacity in feature learning and dimensionality reduction. This paper offers a comprehensive survey, beginning with the fundamentals of conventional autoencoders and their evolution. It categorizes autoencoders by structure and principle, analyzing various models and their applications across domains like machine vision. The study concludes with insights into current limitations and future directions for advancing autoencoder algorithms [31].

III. PROPOSED CONCEPTUAL MODEL

The proposed ensemble model is called RFXG (Random Forest Extreme Gradient) Ensemble, in which Random forests consist of several decision trees. It uses a bagging approach, which includes partitioning the dataset into parallel homogeneous subsets (trees). When constructing each tree, RF uses a random subset of features and a random training data sample to develop a predictive model. The initial level predictions are derived by merging all models/trees and averaging the projected outcomes. The RF method was further improved by assembling it with the GBM. However, the values at each split were examined using all variables, and 200, 350, 500, 750, 1,000, and 2,000 trees were used to test. Figure 1 and 2 shows the working of the proposed RFXG model.

Alternatively, GBM employs boots instead of bags. In boosting, model predictions are enhanced by successively transforming weakly correlated learners (variables) into well-correlated strong learners. Maximum tree depth (interaction depth), number of boosting rounds (or trees), shrinkage, and minimum terminal node size are only some of the aspects of the GBM method that have been modified. RFE, or recursive feature elimination, was the method that was used to determine how important each variable was in the context of the machine-learning algorithms [20]. The RFE makes use of a method called backward selection removal on variables. During this process, a model is fitted using all variables, and the variables that provide the least relevant information are removed after each iteration. They are rebuilding the model and sorting each variable according to its relevance. When applying the RF method to all the data layers in each model included in the training dataset, the RFE performed a 10-fold cross-validation. Following that, the models were sorted according to the normalized RMSE value.

A. DATA COLLECTION

Considering the significance of the proposed research, a parametric dataset comprising primarily year, month, temperature, wind, clouds, and cotton output parameters has been employed for model analysis. Data collection focused on key parameters such as cotton output, year, and land area, resulting in a dataset containing seven columns and 612 rows. These records were sourced from various cities in Pakistan, as previously discussed in the Cotton Production Cities in Pakistan Section 1.5. The Computerized Data Processing Center, the Pakistan Meteorological Department (PMD), and the Crop Reporting Service (CRS) collaboratively designed the dataset with a reference to cotton output. The PMD, an autonomous organization based in Islamabad, Pakistan, is responsible for issuing weather forecasts and warnings for public safety and providing general meteorological information. Beyond meteorology, PMD is engaged in various activities, including crop yield monitoring, astronomical observations, hydrology studies, astrophysics research, climate analysis, and exploration of aeronautical engineering and renewable energy sources across Pakistan. Accessing PMD's data for research purposes involved submitting a signed application from my university supervisor outlining the research's purpose and data importance. After reviewing the application and negotiating terms and conditions, the PMD granted permission to access their data. Table 1 provides an in-depth dataset description, emphasizing its key characteristics and parameters.

Numerous weather parameters are independent and are vital in predicting cotton yield. These parameters significantly impact the production of crops, particularly cotton, because Pakistan has four seasons and, in some areas, cold weather and, in others, warm weather, so research on cotton using weather parameters is crucial. Suppose we can determine which factor influences cotton output most so that we can overcome it, alter the location, or make another production-related choice. In that case, we can boost our production with the help of research. The descriptive statistics of data are shown in Table 2.

B. DATA PREPROCESSING

Data preprocessing plays a crucial role in the generalization performance of supervised algorithms. One significant challenge is noise removal, often involving eliminating instances with excessive null feature values and outliers. Dealing with large datasets often requires selecting representative samples. Managing missing data is another common problem in data preparation, where various uncertainties arise from the reasons behind missing values.

Data Profiling: This initial step involves understanding the dataset's structure and content, identifying the key attributes, and assessing data quality.

Data Reduction: Reducing the volume of data by removing irrelevant or redundant features to streamline analysis without losing significant information.

Data Enrichment: Enhancing the dataset with additional relevant information, possibly by integrating external data sources or creating new derived features.

Data Cleansing: Handling missing values, correcting errors, and removing duplicates to ensure data accuracy and consistency.

*Data Transformation:*Converting data into suitable formats for analysis, such as normalizing scales or encoding categorical variables.

Data Validation: Verifying the processed data's accuracy and completeness to ensure it meets the required quality standards.

Discretization is essential to reduce the number of continuous feature values, but determining interval boundaries and arity for numerical values remains challenging. Discretization methods can be categorized as unsupervised (ignoring class labels) or supervised (considering class labels). Unsupervised methods like equal size and equal frequency aim to simplify feature values without supervision. The process is shown in Figure 3.

The research examined eight factors, including cotton output, wind speed, month, cloud cover, rainfall, and cotton cropland area, with three variables having a climate-related connection. These climate factors influenced cotton production; the dependent variable was measured in bales. Rainfall was recorded in millimeters, temperatures in Celsius, land area in thousands of hectares, wind speed in knots, and cloud cover in oktas. The study utilized two unique datasets without data gaps or duplicates. Still, it adjusted rainfall values below 0.001 millimeters to that threshold for modelling purposes and indicated this change in the dataset. Additionally, data from two different departments were merged, encompassing numerical data spanning from 2005 to 2020, with one dataset focusing on cotton production and area and the other on meteorological characteristics.

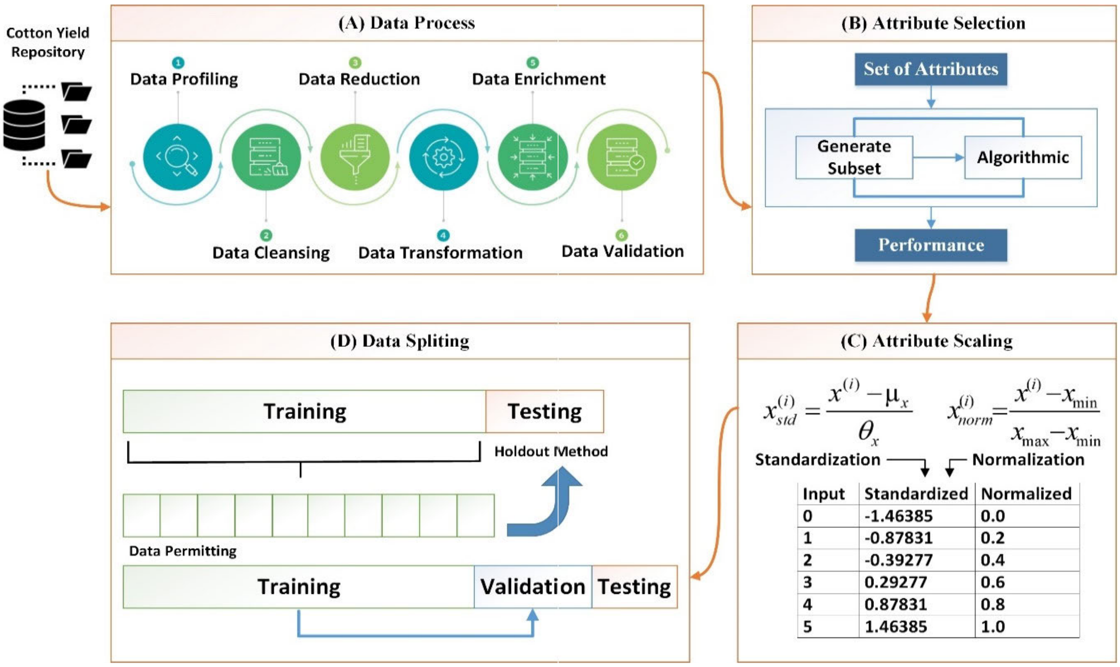
C. ATTRIBUTE AND FEATURE SELECTION

Attributes and Feature selection is the technique of deleting as many unnecessary or duplicate features as feasible from a collection of features, known as subset selection. The feature selection process is shown in Figure 4. The dimensionality of the data is reduced as a result, making it possible for learning algorithms to work in a timelier and more effective manner. In general, these are the kind of features that are described as:

- Relevant: These qualities influence the output and cannot be substituted for others because of their importance.
- Irrelevant: Irrelevant characteristics are characteristics that have no effect on the result and whose values are created randomly for each sample.

Redundant: Redundancy arises when one feature can fulfil the function of another.

FS (Feature Subset) algorithms are divided into two parts: a selection algorithm that generates proposed subsets of features and searches for the optimal subset and an evaluation algorithm that evaluates how 'good' a proposed feature subset is and returns a measure of its goodness to the selection process. The selection algorithm is responsible for creating proposed subsets of features, and the evaluation algorithm is responsible for finding the optimal subset. On the other hand, in the absence of an adequate stopping condition, the FS process has the potential to explore the space of subsets



Input	Standardized	Normalized
0	-1.46385	0.0
1	-0.87831	0.2
2	-0.39277	0.4
3	0.29277	0.6
4	0.87831	0.8
5	1.46385	1.0

FIGURE 1. Proposed RFXG, including data processing and attribute scaling.

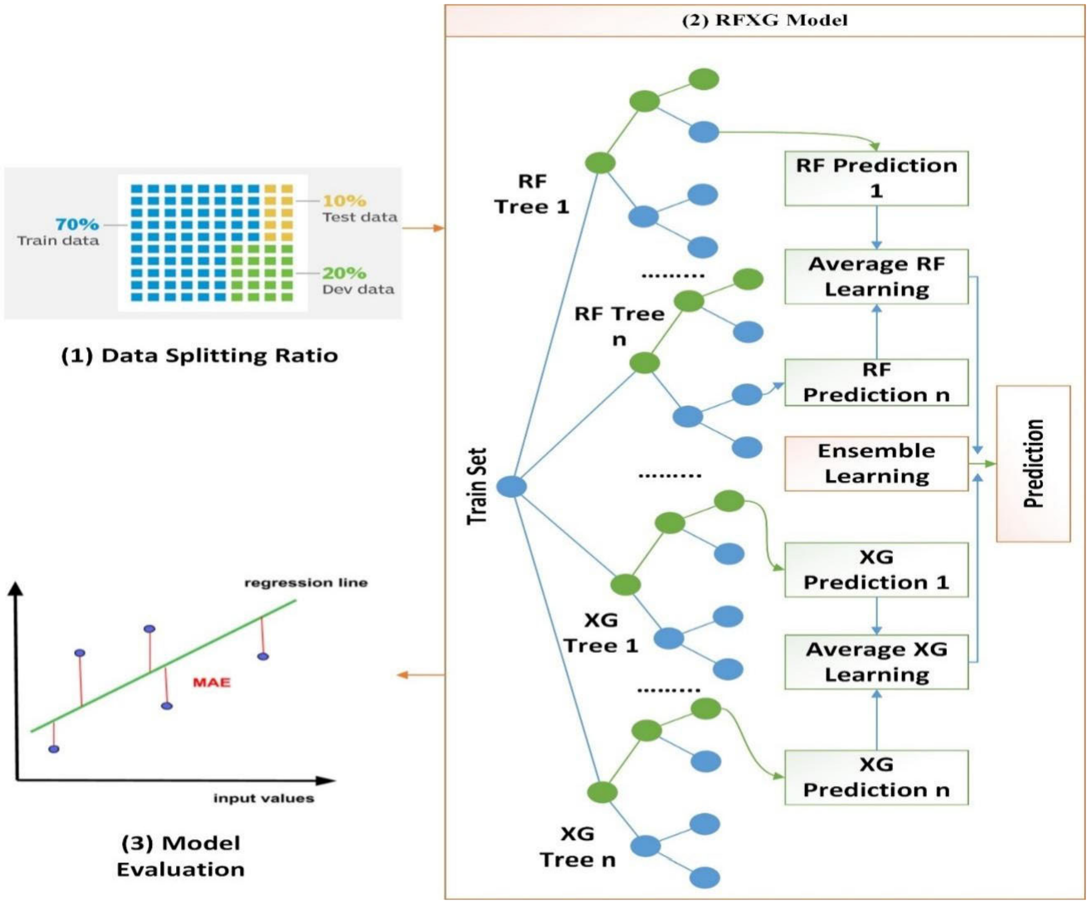


FIGURE 2. Proposed RFXG including prediction and evaluation.

exhaustively or endlessly. The criteria for stopping might include:

(i) If the addition of any characteristic does not result in a more desirable subset

TABLE 1. Dataset description obtained from PMD and CRS.

Years	Months	Temp °C	Wind AT 1200 UTC (Knots)	CLOUD AT 1200 UTC (oktas)	Rain millimeters (mm)	Cotton Production (Bales)
2005	1	13.3	1.3	2.5	64.2	20.54
	2	14.9	2.3	4.3	55.6	20.54
	3	21.6	3	3.4	64.2	20.54
	4	27	4.2	1.6	11.4	20.54
	5	30.7	3.8	2.4	1.7	20.54
	6	34.9	3.3	1.5	32	20.54
	7	30.5	2.5	4.2	223.5	20.54
	8	31.3	3	2.3	129	20.54
	9	29.9	3.3	2.8	69.8	20.54
	10	26.5	1.9	0.2	0.8	20.54
	11	20.5	1.3	0.8	0.01	20.54
	12	14.3	1	1.4	0	20.54

TABLE 2. Descriptive statistics of data.

	Year	Month	Temp °C	Wind AT 1200 UTC (Knots)	Cloud AT 1200 UTC (oktas)	Rain millimeters (mm)	Cotton Production (Bales)
Count	192	192	192	192	192	192	192
Mean	2012.5	6.5	24.6	2.6	2.5	59.3	24.7
Std	4.6	3.4	7.2	2.4	1.3	84.9	10.7
Min	2005	1	0.6	0.2	0	0	10.9
25%	2008.7	3.7	18.5	1.7	1.5	5.3	15.9
50%	2012.5	6.5	26.7	2.5	2.4	22.8	21.1
75%	2016.2	9.2	30.9	3.4	3.5	76.1	32.2
Max	2020	12	35.3	33.4	6.5	450.3	51.6

(ii) Whether or not a subset that is optimum according to some evaluation function is produced.

In an ideal world, feature selection algorithms would search through the various subsets of features to find the contending 2N candidate subsets and then try, based on some evaluation function, to pick the subset that best meets the requirements of the problem at hand. This method is comprehensive since it searches for the best candidate. Even for a feature set of a moderate size, it could be difficult and prohibitively costly to do so (N). Other methods, such as those

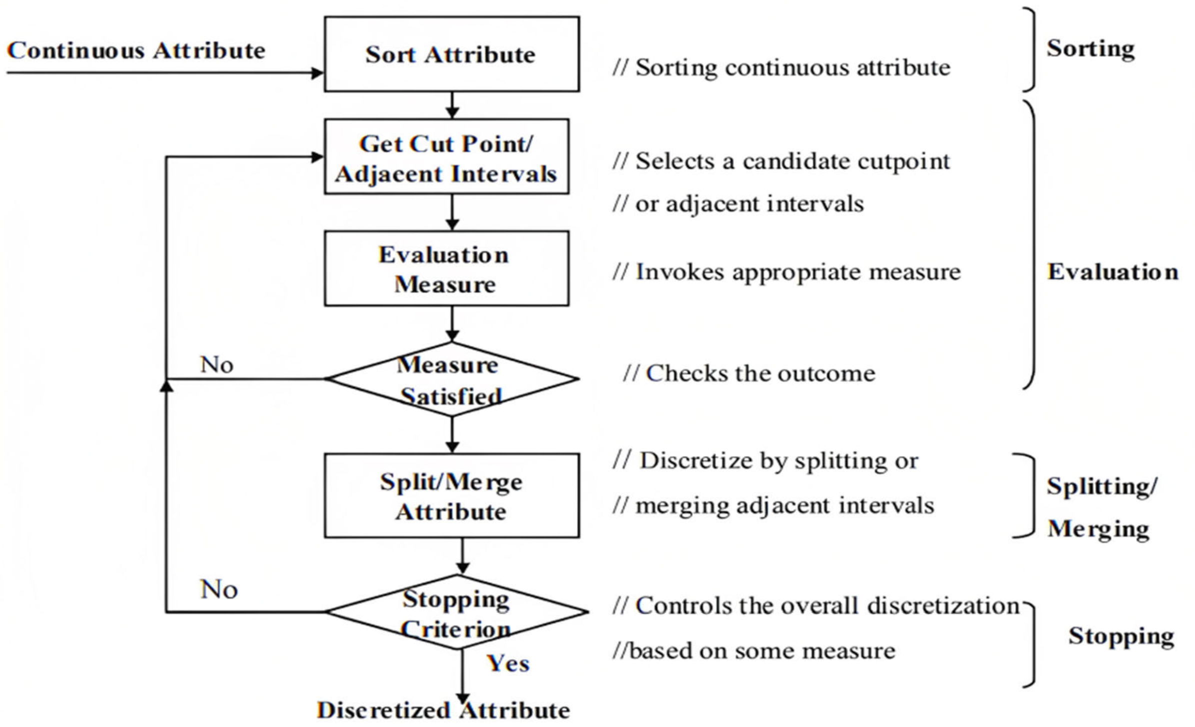


FIGURE 3. Discretizing process adopted from [21].

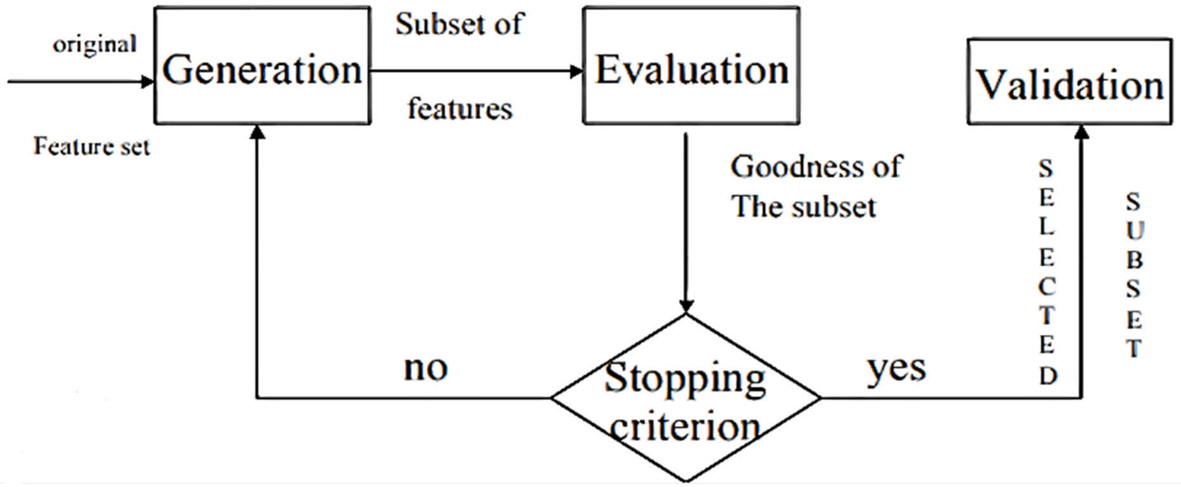


FIGURE 4. Feature subset selection.

based on heuristic or random search strategies, try to reduce computational complexity while maintaining as high a level of performance as possible.

D. ATTRIBUTE AND FEATURE SCALING

With the construction of the feature set, it was noted that various features would not contribute evenly to the model fitting and learning functions, which could result in skewed results. This was one of the factors that led to the formation of the feature set. Before the model fitting step of this study, feature-wise normalization using Min Max Scaling was performed to solve this possible problem. This was done before the model fitting step. No of the unit of measurement, a machine learning algorithm will typically give more weight to higher values and consider smaller values to have less significance.

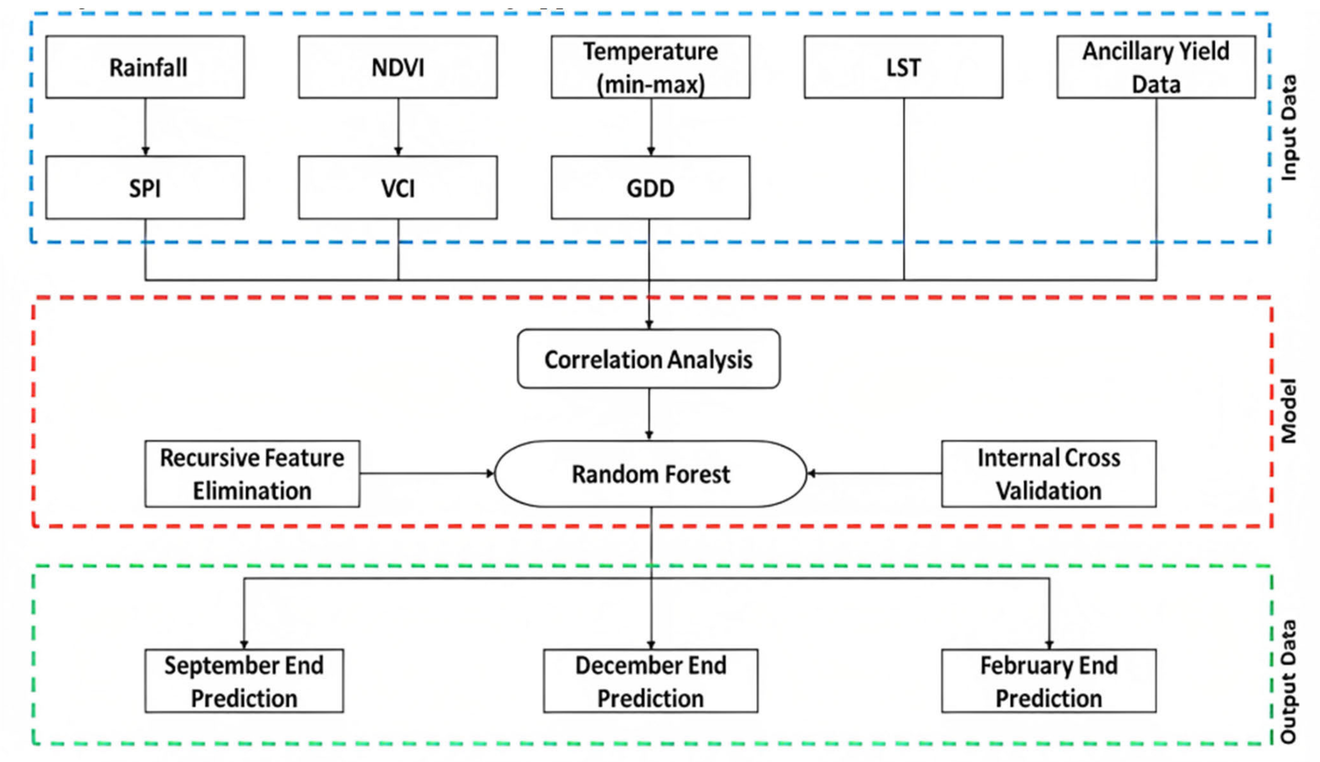


FIGURE 5. Random forests for cotton yield prediction.

This is because conventional approaches do not include any feature scaling.

The issue of various features and attributes may also be solved by utilizing Min-Max Normalization to scale these characteristics [55]. The objective of the min-max normalization is to bring all the numerical values that make up an attribute with numerical values, represented by v, within a predetermined range, denoted by [new $- min_A$, new $- max_A$]. To acquire the new value v, the following equation (equation 1) is applied to the value v to receive the converted value:

$$v = \frac{v - min_A}{max_A + min_A}(new - max_A - new - min_A) + new - min_A \tag{1}$$

The term "normalization" most often refers to a specific instance of the min-max normalization in which the ultimate interval is [1, 0], but there are exceptions to this rule. that is, new $- min_A = 0$ and new $- min_A = 1$. When normalizing the data, it is common practice to use the interval [1, 1].

E. RFXG MODEL-BASED PREDICTION

RFXG model is a hybrid technique consisting of two algorithms named Random Forest and XG-Boost approach. For the more general tasks of classification and regression, RFXG is a machine learning approach that performs very well, can be easily scaled, and is pleasant to users. Specifically, by combining their strengths, RFXG enhances the prediction performance of both Random Forest and XG-Boost. It allows for processing in parallel, which is essential for scalability. Handles missing data by design and adopts the Scikit-Learn API for its user-friendliness. Maintains compatibility with scikit-learn pipelines and model selection tools by adhering to the norms established by scikit-learn.

The RF is based on correlation analysis between the monthly metrology spectral variables and the yearly yield to predict initial level factors that had the greatest impact on yield (see, Figure 6). Here, NDVI is the Normalized Difference Vegetation Index, SPI is the Standardized precipitation index, VCI is the Vegetation condition index, GDD is the Growing degree days, and LST is the Local standard time.

The primary objective of this work was to develop initial level yield estimate models for cotton based on Random Forest (RF)-based modelling inside the R environment with the help of the R package for random Forest. The proposed RF comprises several classification and regression trees (CARTs) that are combined to provide accurate results and prevent the model from being overfitting. Both characteristics are significant in the M_{try} and N_{tree}. M_{try} is the point at which the number of variables that need to be picked and divided at each node equals that number, and N_{tree} is the number of trees that will be planted randomly. A default selection of 500 trees was made; however, the size of the N_{tree} may be increased to any extent conceivable since it does not overfit the model. In most cases, the M_{try} value for regression will be calculated by dividing the total number of observations by three. However, key functionality has been elaborated in the subsequent sections.

- **Normalized Difference Vegetation Index (NDVI) and Vegetation Citation Index (VCI)**

The Landsat Normalized Difference Vegetation Index (NDVI) is a tool for quantifying vegetation greenness and assessing changes in plant health. The Vegetation Condition Index (VCI) evaluates the current NDVI concerning the range of values seen during the same period in preceding years. The VCI, presented in percentages, provides a sense of where the observed value falls about the extreme values (minimum and maximum) from prior years. Lower and higher numbers represent poor and good vegetative state conditions.

- **Standardized precipitation index (SPI)**

The SPI has seen considerable usage as a probability indicator that generally identifies the scenario of extreme dryness and wetness. It has been used across the world to describe weather conditions of drought. The SPI is usually calculated after a substantial amount of time has passed in a particular area. It was discovered that the gamma distribution was extremely appropriate for the SPI precipitation. One would require long-term precipitation data to calculate SPI in a perfect world. This study determined the SPI using CHIRPS from 1981 to 2017 (at 0.05° precision). The average amount of precipitation falling throughout the growing season was calculated for five districts. Because using a formula to calculate SPI is not a realistic option, an SPEI package constructed in the R environment was used to do the computation [56]. The SPI for one month has been computed, and its results have been included in the model. Since the 1-month SPI is a short-term indicator that reflects the crop moisture index over the whole crop growth season, its primary purpose is to investigate crop stress and soil moisture levels.

- **Growing degree days (GDD)**

Increasing degree days, known as GDD, are frequently referred to as heat units. GDD is a measurement that tracks the accumulation of heat units as the crop progresses through its many phenological stages. The flower opens its petals, and the crop matures at this significant unit. The GDD varies depending on the crop. To calculate GDD, the ECMWF provided the lowest and maximum temperatures for each day in NETCDF format (European Center for Medium Weather Forecast) [22].

Alternatively, at the final level, Gradient tree boosting (GTB) has achieved an accurate result in several conventional regression and classification methods. This has been the case in several instances. A regression tree comprises a selected scaled feature one after the other. When applied to a data point, the feature values determine the outcome value, which is determined by the scaled features. Regression trees are built using training data to locate the best possible prediction. The obtained ensemble trees are a family of K regression trees in which the end prediction is the total of the predictions made by the individual trees. The optimal tree ensemble is found by XG boost by repeatedly adding new trees to the existing ensemble in such a way as to maximize the prediction performance of the trees already present in the ensemble as well as the performance of the new tree that is being added. The training data is used to discover the optimal tree ensemble. This is called the "enhanced" version of the traditional XG boost. The functionality of enhanced XG boost is expected to significantly improve over traditional XG boost. It carries out this procedure iteratively until the residuals have been reduced to the utmost degree [23]. Therefore, the gradients of the performance assessment functions are employed in building the new tree to identify the ideal prediction.

IV. EXPERIMENTAL RESULTS

This section presents experimental findings, comparisons, and analyses within the conceptual framework. It assesses the growth phases of cotton plants. It examines the impact of factors such as rainfall, wind, and insect attacks based on field observations and experiments conducted on cotton fields. The results of these experiments and comparisons show that the proposed conceptual framework's performance is comparable to existing approaches. Multiple experiments use diverse datasets to evaluate the suggested framework's effectiveness.

A. PERFORMANCE EVALUATION

The selection of evaluation metrics is contingent on the nature of the issue at hand and the metrics that work well for that problem type. After training the model using the disclosed dataset, we will assess our approach using the test dataset. For the validation of models, hidden values that were not used during training are provided.

1) ROOT MEAN SQUARED ERROR (RMSE)

The root-mean-square error, or RMSE, is a loss function often used in regression applications.

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(yi - \tilde{y}i)2} \tag{2}$$

2) MEAN SQUARED ERROR (MSE)

Finding the squared difference between the actual value and the projected value is what is meant by the term "mean squared error." It is the square of the difference between the observed and predicted values.

$$MSE = \frac{1}{n}\sum (yi - \tilde{y}i)\,2 \tag{3}$$

3) MEAN ABSOLUTE ERROR (MAE)

The mean absolute error measures errors in paired observations of the same phenomenon (MAE). The mean absolute error (MAE) is the sum of all absolute errors divided by the number of samples.

$$MAE = \frac{1}{n}\sum (yi - \tilde{x}i) \tag{4}$$

4) MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

The mean absolute % error (MAPE), or the mean absolute percentage deviation, is a metric that evaluates how well a forecasting approach predicts future outcomes. Another name for the MAPE is the mean absolute percentage deviation (MAPD).

$$MAPE = \frac{100\%}{n} \sum (yi - \tilde{yi}) / yi \tag{5}$$

B. BASELINE MODELS

The baseline approaches used for comparisons are given below:

- SVM [24]: A machine learning-based data pipeline accurately predicts cotton breeding field yields from aerial photos, achieving an RMSE of 0.29 compared to ground truth counts.
- RF [25]: Machine learning models effectively predict within-field cotton lint output at critical development phases using publicly available data, with RMSE at 0.63.
- DCNN [26]: a modified DCNN model that outperforms Seg-Net, SVM, and RF in predicting cotton yield from aerial images.
- ANN [27]: The model combines various data, including precipitation, heat units, and vegetation indices, to forecast cotton production four months in advance with an RMSE of 0.52, showing potential for yield prediction in data-limited regions.

C. RESULTS

A training dataset and a testing dataset were created from the available data. Approximately 70% of the datasets were used to train the model, and the remaining 30% of the datasets, which were not utilized to train the model, were used for validation. Utilizing the variable significance graph from RF, the most influential factors on cotton yield were chosen, and the least influential ones were eliminated. The decision was made based on the percentage increase in the value of the mean square error of predicted outcomes of variables being permuted, as well as the loss of node purity for a particular variable across all trees, represented as Inc-Node-Purity. The model would be more accurate and quicker in predicting outcomes if the most influential factors were appropriately selected. In conjunction with internal cross-validation, recursive feature elimination (RFE) was used to minimize the number of variables. The selected set of variables was employed in developing the RF-based model for estimating cotton yield based on the predictor variable significance graph created from the data. The created models were checked for accuracy using the left-behind data.

1) INDIVIDUAL MODELS EVALUATION

This section discusses the performance of the existing cotton prediction models on prescribed datasets. The results of actual prediction that has already been identified by field AARI, WMO and Food and Agriculture Organization (FAO) have been compared with the prediction of existing models and are elaborated in the below subsections.

Figure 6 shows a considerable disparity between the anticipated output of the model and the actual output, thereby inferring that the model's prognostication on measurement accuracy is subpar. It is noteworthy that decision tree regression models tend to overfit when dealing with datasets featuring many features; hence, in the present scenario, the data spans sixteen years, and the decision tree model fails to provide an accurate outcome. Notably, the pliability of the decision tree method renders its outcomes far superior to those obtained via linear regression. The researchers meticulously assessed various factors, including air temperature, precipitation, and evapotranspiration, among other variables. During their investigation, they employed a range of diverse algorithms; nevertheless, the decision tree model yielded an 8.2 RMSE, indicative of a relatively low degree of accuracy.

Determining a dependent variable's value can be achieved using a statistical technique called linear regression. This method considers the relationship between the above dependent and independent variables. The correlation between two variables can be established using linear regression. It is a modelling approach that forecasts the dependent variable's value based on one or more independent variables. The present study used multiple independent variables to predict favorable outcomes. However, as indicated by Figure 7, there exists a disparity between the actual and predicted values, thereby highlighting the limited efficacy of linear regression in accurately forecasting outcomes. The model's root mean squared error was calculated to be 9.71, while the mean squared error was 94.28. The inflexibility of linear regression can lead to underfitting, which may contribute to the elevated error score. These findings are elaborated upon in Figure 7.

Choice trees and linear regression models demonstrate superior accuracy by minimizing the disparity between expected and actual outputs. Random forests, while capable of managing large datasets and providing reliable, easily interpretable forecasts, outperform decision tree methodologies regarding outcome perception. This robustness in random forests stems from its ability to determine outcomes by calculating the average of predictions from multiple decision trees, with potential accuracy improvements achievable by increasing the number of trees. Although it demonstrates a root mean squared error of 0.66 and a mean squared error of 0.43, a random forest analysis study yielded a notably higher RMSE of 10.2 with fewer variables and 4.2 when examining Brazilian air temperature, precipitation, and evapotranspiration. The introduction of supplementary variables and the bagging approach notably enhanced the model's effectiveness.

XGB uses criteria like maximum tree depth to improve. XGB Boosting improves accuracy. Once each model is fitted,

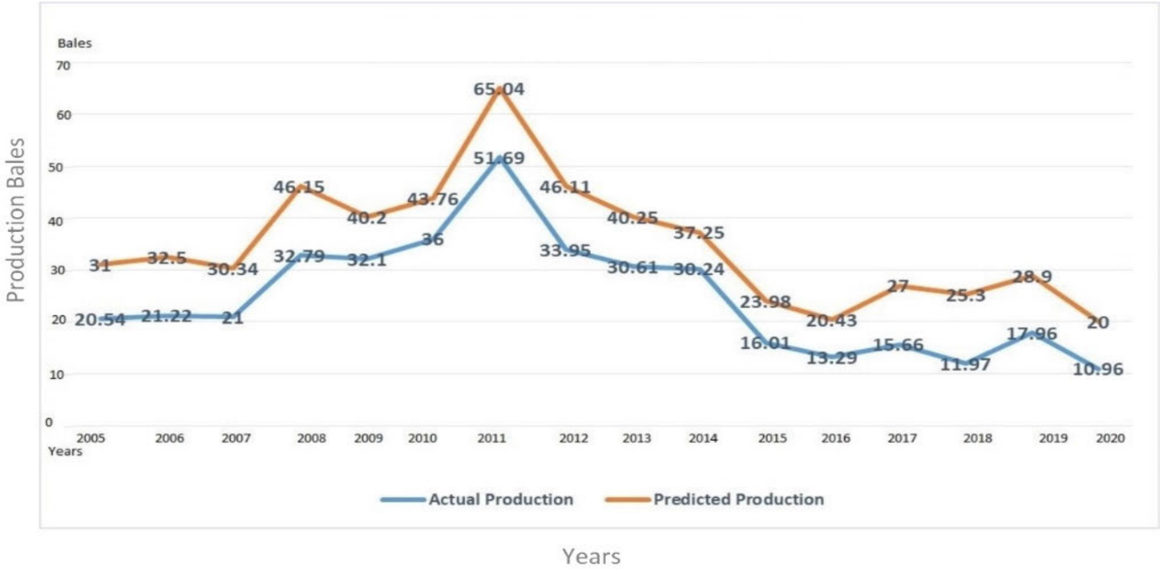


FIGURE 6. Percentage comparison of actual and forecast production using decision tree.

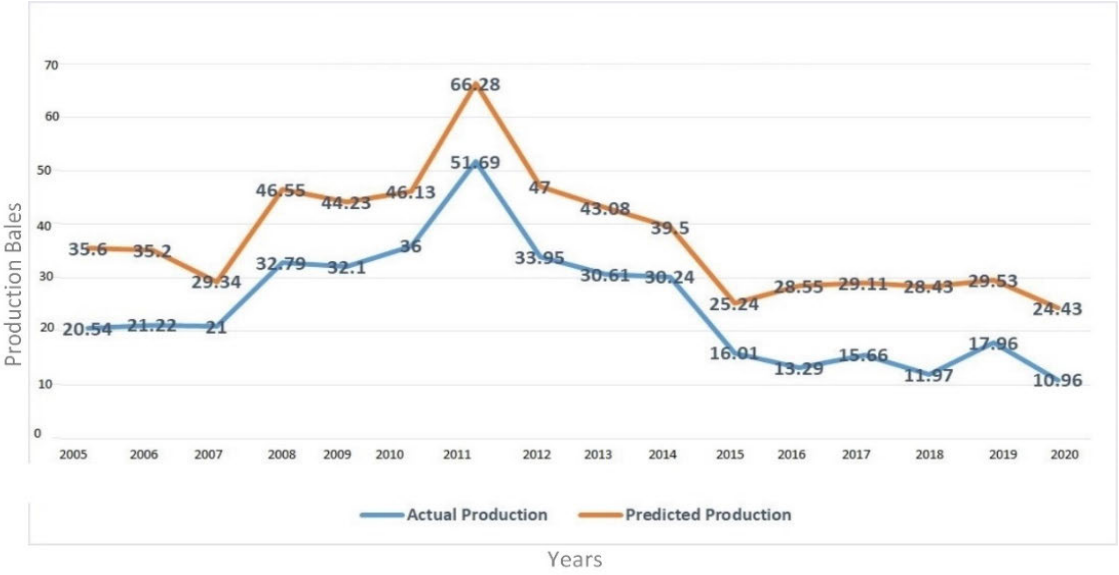


FIGURE 7. Percentage comparison of actual and forecast production using regression.

it is added to the ensemble and attempts to correct the predictions of the previous model. Ensemble learning combines numerous purpose-built weak learners to create a strong learner. Noise, variance, and bias are the three forms of errors that are likely to occur most often in learning models. The ensemble techniques used in machine learning lessen the influence of these error-causing factors, contributing to the accuracy and stability of the machine learning (ML) algorithms. After automatically attempting various parameter combinations, I use cross-validation to choose the best choice to go on to the next step. This takes a longer process, resulting in an accuracy of 0.07 on the tested sample (MSE). In XGB, weight is an extremely essential factor. The temperature, wind speed, cloud cover, and rainfall are the four independent variables that have weights given to them before input into the decision tree that forecasts the outcomes. The XGB model, which has an RMSE of 0.27 and an MSE of 0.07, achieves the greatest results across all climatic factors, as shown in Figure 9.

Compared to other models, an efficient one will be able to make more accurate predictions of the testing data and, hence, will be the one that is ultimately implemented effectively.

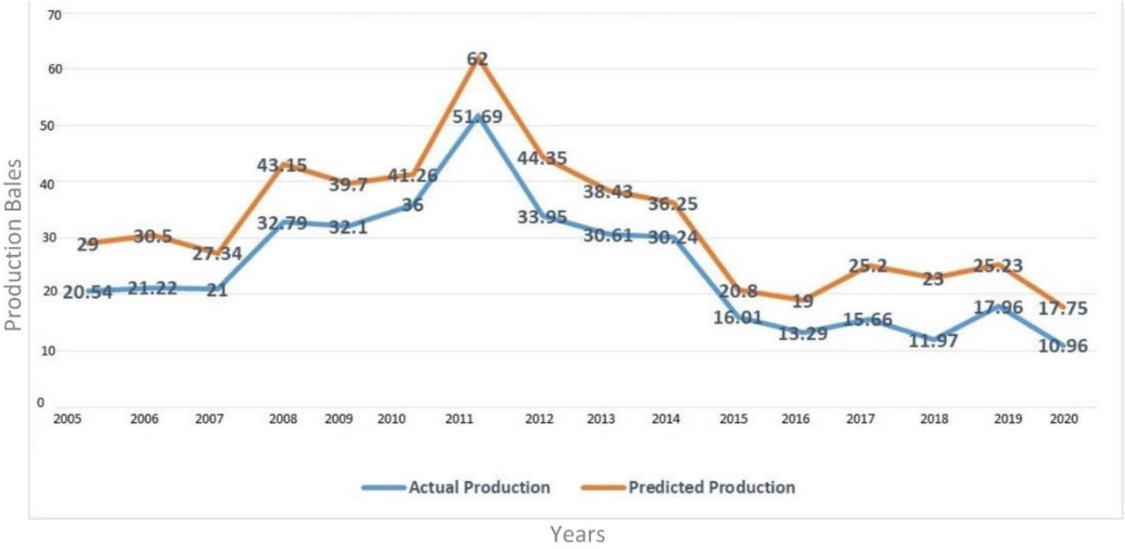


FIGURE 8. Percentage comparison of actual and forecast production using random forest.

Figure 10 shows the performance of the proposed RFXG with the observed data. The yearly prediction and its numerical values depicted the performance of the proposed work on the same dataset as tremendous.

2) PROPOSED MODEL RESULTS OF FIVE DIFFERENT CITIES

Accurately monitoring and assessing crop yields is of utmost significance for various cities in Pakistan. This critical task holds tremendous importance because crop yield, particularly cotton, is the most influential agricultural factor for sustainable development. Determining crop yield holds immense value in formulating policies that cater to the nation's economic interests. Consequently, evaluating the potential crop yields at an early stage becomes essential for making informed agricultural decisions and directly impacts the financial outcomes for farmers. Figure 11 presents the recorded outcomes for five cities that have consistently produced substantial cotton yields over 15 years, from 2005 to 2019.

D. COMPARISON OF RFXG WITH BASELINES APPROACHES

The indices of the proposed RFXG approach and the baseline method are shown in Table 3. It has been shown that the RFXG is an innovative model that yields very good outcomes compared to the approaches used in the baselines. The baseline approaches are:

TABLE 3. Comparisons of proposed RFXG with baselines approaches.

Approaches	RMSE	MSE	MAE	MAPE
SVM	0.29	0.09	1.56	93%
RF	0.63	1.35	2.03	88%
DCNN	0.31	0.10	1.72	94%
ANN	0.52	0.12	2.01	89%
Proposed Approach	**0.22**	**0.05**	**1.23**	**86%**

The random forest technique combines the outputs of multiple decision trees to arrive at a final prediction, which helps to increase the accuracy of the forecast. Adjusting the training data is yet another technique that can be used to enhance the trained models. In ensemble-based approaches, each consecutive model strives to rectify the predictions provided by the previous model. This iterative process helps to improve the accuracy of the model over time. Machine learning models are susceptible to errors caused by noise, volatility, and bias. Ensemble-based approaches effectively mitigate these errors by aggregating the predictions of multiple models, thereby reducing the variance of the model and improving its reliability. In summary, the proposed ensemble-based RFXG model is a powerful tool that can help to improve the accuracy and reliability of machine learning models used in various applications. Random Forest Extreme Gradient Boosting (RFXG) combines the strengths of Random Forests and gradient boosting, offering a robust approach to feature selection and model performance. Random Forests excel in feature selection by evaluating feature importance across numerous trees, thereby reducing overfitting and enhancing generalization. This robustness is enhanced by optimization that adapts to the needs of each iteration, leading to optimal use of the. The flexibility of the function allows RFXG to continuously improve useful features, even in the presence of high-dimensional or noisy data. In contrast, XGBoost builds

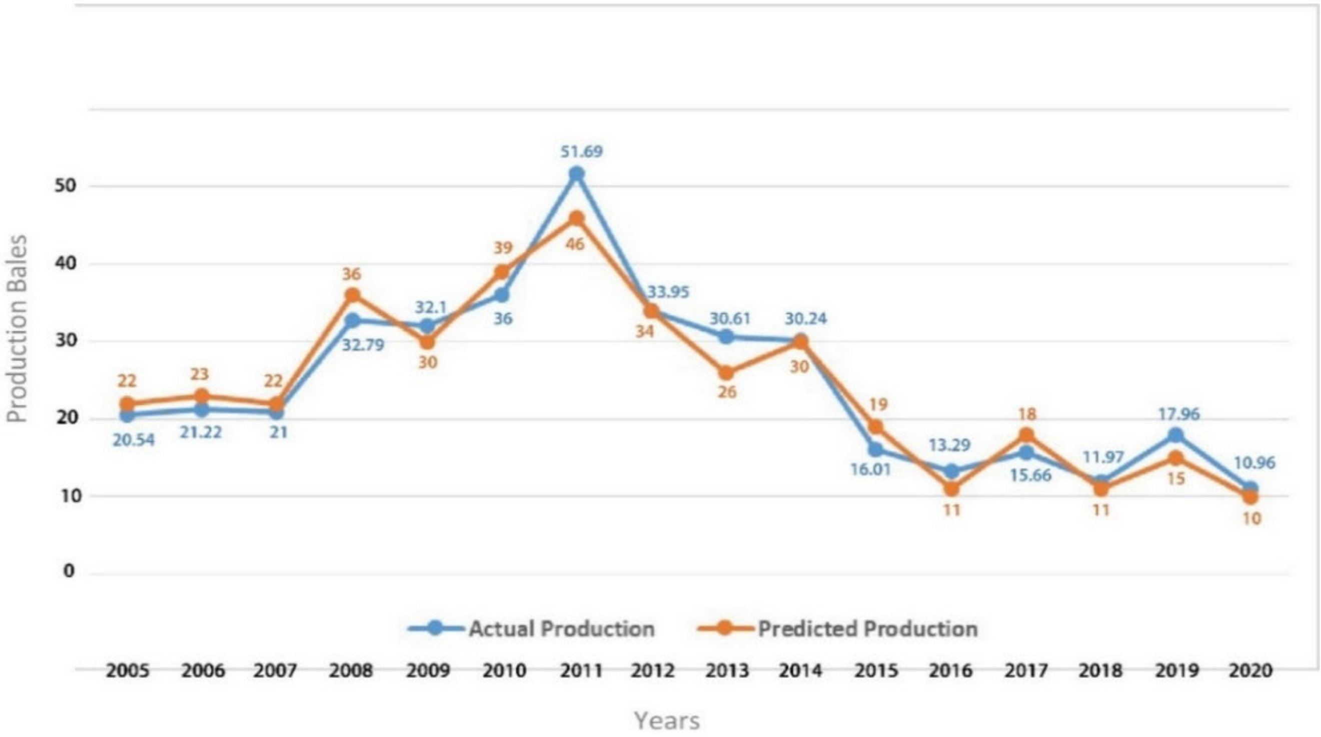


FIGURE 9. Percentage comparison of actual and forecast production using random forest.

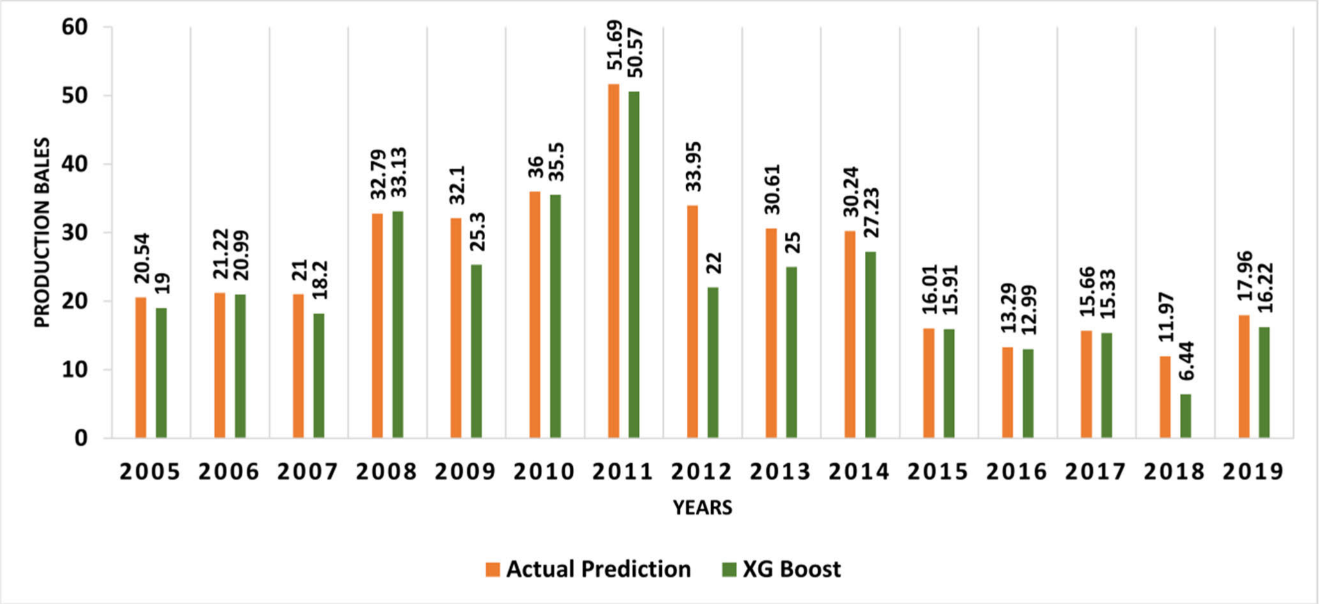


FIGURE 10. Percentage comparison of actual and forecast production using RXFG.

trees sequentially, which can sometimes be less effective in identifying and leveraging relevant features due to its single tree-building approach. While XGBoost provides feature importance metrics through gain, cover, and frequency, these metrics may not capture the holistic importance of features as effectively as RFXG's ensemble method. RFXG's ability to combine robust feature selection with iterative refinement results in superior model performance and generalization. This makes RFXG particularly advantageous in scenarios with complex feature interactions and noisy datasets, offering

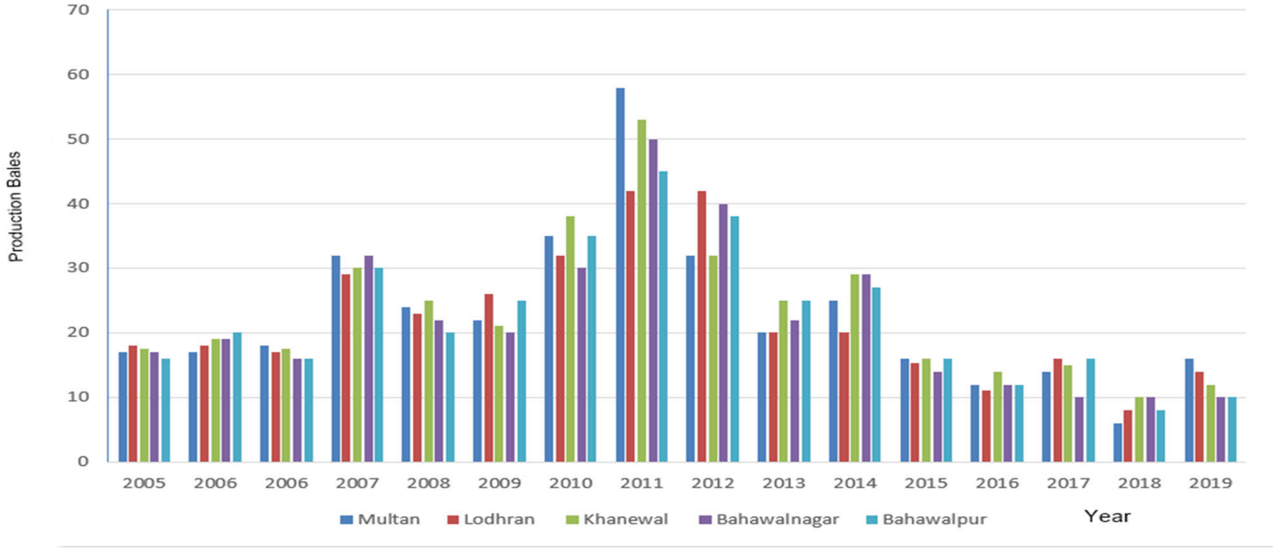


FIGURE 11. **Percentage comparison of proposed techniques of multiple cities.**

a more balanced and effective approach compared to traditional XGBoost.

V. CONCLUSION

The study underscores the critical importance of employing machine learning techniques for accurate and timely predictions of cotton production, benefiting farmers, policymakers, and various stakeholders in the cotton production process. For farmers, precise yield projections inform planting, harvesting, marketing, and risk management decisions. Policymakers can make informed choices regarding agricultural policies and regional interventions, while textile manufacturers and consumers benefit from better planning and product availability. The research introduces an innovative RFXG model, informed by extensive weather parameter analysis and field observations, demonstrating its superiority to existing methods. The study's focus on Pakistan's diverse climate underscores the model's significance. The subsequent chapters detail the model's conceptual framework, data processing phases, and experimental results, showcasing its outperformance of traditional methods. Combining random forest and XGBoost models through the ensemble based RFXG model offers speed, efficiency, and enhanced accuracy by leveraging techniques like bagging, stacking, and boosting. Through iterative model refinement, this approach effectively mitigates noise, volatility, and bias, making the proposed ensemble-based RFXG model a powerful tool for improving the precision and reliability of machine-learning models across various applications.

The implementation of advanced predictive models like RFXG can significantly enhance cotton production practices and policy-making in Pakistan. The model's ability to provide accurate and reliable yield predictions empowers both policymakers and farmers to make more informed, data-driven decisions. This leads to improved resource allocation, (ensuring that areas with higher predicted yields receive the necessary support to maximize output) risk management, financial planning, and market strategies, ultimately enhancing productivity, sustainability, and economic outcomes in the agricultural sector.

In future work, a vast array of alternative methods can be explored for improvement, including the application of suggested approaches to daily weather factors used in forecasting and weather-related health issues. Investigating scalability solutions to adapt RFXG for big data applications. This includes exploring cloud-based platforms, distributed machine learning frameworks, and incremental learning techniques to handle large-scale data more effectively Implementing automated hyper parameter tuning methods such as Bayesian optimization, genetic algorithms, or reinforcement learning to streamline the process and find optimal configurations more efficiently. it is anticipated that future research will increasingly consider the influence of various meteorological indices such as soil temperature, CO2 levels, humidity, radiation levels, and more. Additionally, factors like soil quality, irrigation practices, and pest management could improve the robustness of the predictive model. which are likely to have a more pronounced impact on the outcomes.

REFERENCES

[1] Z. Li, P. Yang, H. Tang, W. Wu, H. Yin, Z. Liu, and L. Zhang, "Response of maize phenology to climate warming in Northeast China between 1990 and 2012," *Regional Environ. Change*, vol. 14, no. 1, pp. 39–48, Feb. 2014.

[2] M. A. Imran, A. Ali, M. Ashfaq, S. Hassan, R. Culas, and C. Ma, "Impact of climate smart agriculture (CSA) practices on cotton production and livelihood of farmers in Punjab, Pakistan," *Sustainability*, vol. 10, no. 6, p. 2101, Jun. 2018.

[3] F. Zulfiqar, A. Datta, and G. B. Thapa, "GB determinants and resource use efficiency of 'better cotton': An innovative cleaner production alternative," *J. Clean. Prod.*, vol. 66, pp. 1372–1380, Nov. 2017.

[4] D. R. Cutler, T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, Nov. 2007, doi: 10.1890/07-0539.1.

[5] M. Rashid, Z. Husnain, U. Shakoor, and M. I. U. Husnain, "Impact of climate change on cotton production in pakistan: An ARDL bound testing approach," *Sarhad J. Agricult.*, vol. 35, no. 1, pp. 333–341, 2020.

[6] P. N. J. V. Singh and M. A. A. Tushar Rathod, "Trend analysis of monthly and annual rainfall in different tahsils of Ahmednagar district," *Int. J. Current Microbiol. Appl. Sci.*, vol. 10, no. 8, pp. 194–209, Aug. 2021.

[7] W. Schlenker and M. J. Roberts, "Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 37, pp. 15594–15598, 2009, doi: 10.1073/PNAS.0906865106.

[8] F. Alam, M. Salam, N. A. Khalil, O. Khan, and M. Khan, "Rainfall trend analysis and weather forecast accuracy in selected parts of Khyber Pakhtunkhwa, Pakistan," *Social Netw. Appl. Sci.*, vol. 3, no. 5, p. 575, May 2021, doi: 10.1007/S42452-021-04457-Z.

[9] A. K. Patil, "Design and development of a robust deep learning model for detection of disease cotton crop," *J. Positive School Psychol.*, vol. 6, no. 3, pp. 6227–6233, 2022.

[10] M. Sharma, N. Mittal, A. Mishra, and A. Gupta, "Machine learning-based electricity load forecast for the agriculture sector," *Int. J. Softw. Innov.*, vol. 11, no. 1, pp. 1–21, Dec. 2022.

[11] A. Samak, S. Rajeswari, R. Ravikesavan, S. Lokanadhan, N. Ganapathy, and A. V. Kini, "Effect of environment on combining ability, heterosis for seed cotton yield in bt introgressed lines of upland cotton," *Agronomy J.*, vol. 114, no. 2, pp. 935–947, Mar. 2022.

[12] N. Mikail and M. F. Baran, "Application of artificial intelligence methods to predict cotton production in Turkey," *Türk Tarim ve Doga Bilimleri Dergisi*, vol. 8, no. 4, pp. 1018–1027, 2021.

[13] R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. V. Prasad, and I. A. Ciampitti, "Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in Southern Brazil," *Agricult. Forest Meteorol.*, vol. 284, Apr. 2020, Art. no. 107886.

[14] L. Parviz, "Assessing accuracy of barley yield forecasting with integration of climate variables and support vector regression," *Annales Universitatis Mariae Curie-Sklodowska, sectio C-Biologia*, vol. 73, no. 1, pp. 19–30, Jun. 2019.

[15] N. Gandhi, O. Petkar, and L. J. Armstrong, "Rice crop yield prediction using artificial neural networks," in *Proc. IEEE Technol. Innov. ICT Agricult. Rural Develop. (TIAR)*, Chennai, India, Jul. 2016, pp. 105–110.

[16] J. Han, Z. Zhang, J. Cao, Y. Luo, L. Zhang, Z. Li, and J. Zhang, "Prediction of winter wheat yield based on multi-source data and machine learning in China," *Remote Sens.*, vol. 12, no. 2, p. 236, Jan. 2020.

[17] Y. Lv, Q.-T. Le, H.-B. Bui, X.-N. Bui, H. Nguyen, T. Nguyen-Thoi, J. Dou, and X. Song, "A comparative study of different machine learning algorithms in predicting the content of ilmenite in titanium placer," *Appl. Sci.*, vol. 10, no. 2, p. 635, Jan. 2020, doi: 10.3390/APP10020635.

[18] L. E. de Oliveira Aparecido, G. de Souza Rolim, J. R. da Silva Cabral De Moraes, C. T. S. Costa, and P. S. de Souza, "Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases," *Int. J. Biometeorol.*, vol. 64, no. 4, pp. 671–688, Apr. 2020.

[19] D. Sotto, A. Philippi, T. Yigitcanlar, and M. Kamruzzaman, "Aligning urban policy with climate action in the global south: Are Brazilian cities considering climate emergency in local planning practice?" *Energies*, vol. 12, no. 18, p. 3418, Sep. 2019.

[20] A. Nawaz, A. A. Awan, T. Ali, and M. R. R. Rana, "Product's behaviour recommendations using free text: An aspect based sentiment analysis approach," *Cluster Comput.*, vol. 23, no. 2, pp. 1267–1279, Jun. 2020.

[21] R. P. L. Durgabai, P. Bhargavi, and S. Jyothi, "Classification of cotton crop pests using big data analytics," in *Proc. Int. Conf. Comput. Bio Eng.* Cham, Switzerland: Springer, Dec. 2019, pp. 37–45.

[22] A. Imran, "Cotton crop development in Faisalabad Central Punjab," Dept. Ayub Agricult. Res. Inst., Regional Agromet Centre Pakistan Meteorological, Faisalabad, Pakistan, Tech. Rep., 2019. [Online]. Available: https://namc.pmd.gov.pk/assets/crop-reports/1342748710Crop-report-Cotton-Faisalabad-2019.pdf

[23] D. H. Kazmi, J. Li, C. Ruan, S. Zhao, and Y. Li, "A statistical downscaling model for summer rainfall over Pakistan," *Climate Dyn.*, vol. 47, nos. 7–8, pp. 2653–2666, Oct. 2016, doi: 10.1007/S00382-016-2990-1.

[24] H. Zhang, P. Wu, A. Yin, X. Yang, M. Zhang, and C. Gao, "Prediction of soil organic carbon in an intensively managed reclamation zone of Eastern China: A comparison of multiple linear regressions and the random forest model," *Sci. Total Environ.*, vol. 592, pp. 704–713, Aug. 2017.

[25] C. Dang, Y. Liu, H. Yue, J. Qian, and R. Zhu, "Autumn crop yield prediction using data-driven approaches: Support vector machines, random forest, and deep neural network methods," *Can. J. Remote Sens.*, vol. 47, no. 2, pp. 162–181, Mar. 2021.

[26] J. Rodriguez-Sanchez, C. Li, and A. H. Paterson, "Cotton yield estimation from aerial imagery using machine learning approaches," *Frontiers Plant Sci.*, vol. 13, Apr. 2022, Art. no. 870181.

[27] S. Leo, M. D. A. Migliorati, and P. R. Grace, "Predicting within-field cotton yields using publicly available datasets and machine learning," *Agronomy J.*, vol. 113, no. 2, pp. 1150–1163, Mar. 2021.

[28] S. Abdollahi, A. Deldari, H. Asadi, A. Montazerolghaem, and S. M. Mazinani, "Flow-aware forwarding in SDN datacenters using a knapsack-PSO-based solution," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 3, pp. 2902–2914, Sep. 2021, doi: 10.1109/TNSM.2021.3064974.

[29] S. Tripathy and M. Tabasum, "Autoencoder: An unsupervised deep learning approach," in *Emerging Technologies in Data Mining and Information Security*, vol. 490, P. Dutta, S. Chakrabarti, A. Bhattacharya, S. Dutta, and C. Shahnaz, Eds., Singapore: Springer, 2023, doi: 10.1007/978-981-19-4052-1_27.

[30] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: A survey," *Artif. Intell. Rev.*, vol. 57, no. 2, p. 28, Feb. 2024.

[31] S. Abdollahi, A. Deldari, H. Asadi, A. Montazerolghaem, and S. M. Mazinani, "Flow-aware forwarding in SDN datacenters using a knapsack-PSO-based solution," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 3, pp. 2902–2914, Sep. 2021.

SYED TAHSEEN HAIDER received the Bachelor of Science degree (Hons.) in computer science from the Institute of Computing and Information Technology, Kohat University of Science and Technology, Kohat, Pakistan, in 2016, and the Master of Science (M.S.) degree in computer science, in 2023. Currently, he is a Lecturer with the University Institute of Information Technology (UIIT), PMAS Arid Agriculture University, Rawalpindi, Pakistan. With over seven years of industry experience, he has expertise in Web Technologies, Java, ASP.Net, C#, Python, Databases, and Oracle. His research interests include data mining, machine learning, social media analysis, medical image diagnosis, graph mining, and big data analytics.

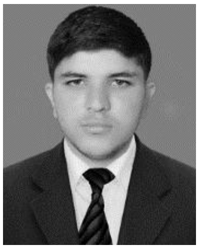
WENPING GE received the B.E. degree in optics from Sichuan University, Chengdu, China, in 1989, the M.S. degree in optical engineering from Xi'an Institute of Optics and Precision Mechanics, Xi'an, China, in 2000, and the Ph.D. degree in electromagnetic field and microwave technology from Shanghai Jiaotong University, Shanghai, China, in 2003. She has been a Faculty Member with Xinjiang University, since 2003, where she is currently a Professor. Her interests include mobile communication, optical fiber communication, and fiber technology.

JIANQIANG LI (Senior Member, IEEE) received the B.S. degree in mechatronics from Beijing Institute of Technology, Beijing, China, in 1996, and the M.S. and Ph.D. degrees in control science and engineering from Tsinghua University, Beijing, in 2001 and 2004, respectively. He was a Researcher with the Digital Enterprise Research Institute, National University of Ireland, Galway, from 2004 to 2005. From 2005 to 2013, he was with the NEC Laboratories China, as a Researcher. He was with the Department of Computer Science, Stanford University, as a Visiting Scholar, from 2009 to 2010. He joined Beijing University of Technology, Beijing, in 2013, as a Beijing Distinguished Professor. His research interests include information retrieval, privacy protection, and big data. He was a PC member of the multiple international conferences and organized the IEEE workshop on medical computing. He served as a Guest Editor to organize a Special Issue on Information Technology for Enhanced Healthcare Services.

SAIF UR REHMAN received the M.C.S. degree (Hons.) from the Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan, Pakistan, in 2005, and the M.S. and Ph.D. degrees in computer science, in 2010 and 2019, respectively. He is currently an Assistant Professor with the University Institute of Information Technology (UIIT), PMAS Arid Agriculture University, Rawalpindi, Pakistan. He has more than eight years of industry experience and involved in Java, ASP.Net, C#, and Oracle. He is also an advisor in PPSC and an Examiner in PPSC and KP-PSC. He has published five book chapters and more than 60 research articles in different renowned international Q1 and Q2 journals. His research interests include data mining, graph mining, social graph analysis, and big data analytics.

AZHAR IMRAN (Senior Member, IEEE) received the master's degree in computer science from the University of Sargodha, Pakistan, and the Ph.D. degree in software engineering from Beijing University of Technology, China. He is also an Assistant Professor with the Department of Creative Technologies, Faculty of Computing & Artificial Intelligence, Air University, Islamabad, Pakistan. He was a Senior Lecturer with the Department of Computer Science, University of Sargodha, Pakistan, from 2012 to 2017. He is also a Renowned Expert in image processing, healthcare informatics, and social media analysis. He has over 12 years of national and international academic experience as a full-time Faculty Member. His research interests include image processing, social media analysis, medical image diagnosis, machine learning, and data mining. He aims to contribute to interdisciplinary research of computer science and human-related disciplines. He is a Regular Member of IEEE and has contributed with more than 60 research articles.

MOHAMED ABDEL FATTAH SHARAF received the Ph.D. degree in industrial engineering from Chiba University, Japan. He is currently the Head of the Development and Quality Unit, College of Engineering, King Saud University. He has published more than 30 articles in the areas of spare parts control, quality management, maintenance, six sigma methodology, and academic accreditation.



SYED MUHAMMAD HAIDER received the B.S. degree in computer science from the University Institute of Information Technology (UIIT), PMAS Arid Agriculture University, Rawalpindi, Pakistan, in 2019, and the M.S. degree in computer science from Abasyn University Islamabad Campus, in 2024. His research interests include data mining, graph mining, social graph analysis, and big data analytics.