IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 February 2024, accepted 10 March 2024, date of publication 18 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3376735

RESEARCH ARTICLE

Accurate Wheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion

MUHAMMAD ASHFAQ[1], IMRAN KHAN[2], ABDULRAHMAN ALZAHRANI[3], MUHAMMAD USMAN TARIQ[4], (Member, IEEE), HUMERA KHAN[5], AND ANWAR GHANI[2,6]

[1]Department of Software Engineering, International Islamic University Islamabad, Islamabad 44000, Pakistan
[2]Department of Computer Science, International Islamic University Islamabad, Islamabad 44000, Pakistan
[3]Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia
[4]Marketing, Operations, and Information System, Abu Dhabi University, Abu Dhabi, United Arab Emirates
[5]Department of Information Systems, Faculty of Computing and Information Technology, Northern Border University, Rafha 76312, Saudi Arabia
[6]Big Data Research Center, Jeju National University, Jeju-si, Jeju-do 63243, South Korea

Corresponding author: Anwar Ghani (anwar.ghani@iiu.edu.pk)

ABSTRACT Due to exponential population growth, climate change, and an increasing demand for food, there is an unprecedented need for a timely, precise, and dependable assessment of crop yield on a large scale. Wheat, a staple crop worldwide, requires accurate and prompt prediction of its output for global food security. Traditionally, the development of empirical models for crop yield forecasting has relied on climate data, satellite data, or a combination of both. Despite the enhanced performance achieved by integrating satellite and climate data, the contributions from various sources (Climate, Soil, Socioeconomic, and Remote sensing) remain unclear. The lack of well-defined comparisons between the performance of regression-based approaches and different Machine Learning (ML) methods in yield prediction necessitates further investigation. This study addresses the gaps by combining data from multiple sources to forecast wheat yield in the Multan region in the Punjab province of Pakistan. The findings are compared to the benchmark provided by Crop Report Services (CRS) Punjab, with three widely used ML techniques (support vector machine (SVM), Random Forest (RF), and Least Absolute Shrinkage and Selection Operator (LASSO)) by integrating publicly available data within the GEE (Google Earth Engine) platform, including climate, satellite, soil properties, and spatial information data to develop alternative empirical models for yield prediction using data from 2017 to 2022, selecting the best attribute subset related to crop output. The district-level simulated yield data set was analyzed with three ML models (SVM, RF, and LASSO) as a function of seasonal weather, satellite, and soil. The results indicate that combining all datasets using three ML algorithms achieves better yield prediction performance (R^2: 0.74-0.88). Incorporating spatial information and other properties into benchmark models can improve the prediction from 0.08 to 0.12. Random forest outperformed the competitor models with a Root Mean Square Error (RMSE) of 0.05 q/ha and R^2 of 0.88. Comparative analysis shows that random forest with 97% and SVM with 93% yielded better results in the study area.

INDEX TERMS Machine learning, RF, LASSO, remote sensing, SVM, CNN, crop yield prediction.

I. INTRODUCTION

Wheat, one of the three principal crops farmed globally, is a substantial source of calories, protein, and vital micronutrients for people [1], [2]. Wheat continued high-yield potential, on the other hand, faces significant threats due to a variety of production restrictions, including rising temperatures, increased precipitation unpredictability, and frequent extreme weather events [3], [4]. Accurate crop output forecasts before harvest, thus ensuring food security and trade. Traditional agricultural yield evaluation methods include conducting field surveys during the growing season or relying on prior

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

knowledge of the crop-growing environment. Conventional yield estimates, however, are difficult due to problems such as small sample sizes, insufficient staff for required sampling frequency and size, and the impact of inter-annual climate variability. Data processing concerns such as sampling and non-sampling mistakes can also cause dependability issues and inaccuracies [5].

ML techniques are increasingly used in various fields like health [6] and education [7]. To address these challenges, researchers increasingly turn to nonlinear models for agricultural yield estimation [1], [8]. The application of ML for estimating agricultural yields has gained significant attention in recent years [9], [10]. Various ML techniques have been extensively utilized to achieve precise predictions for the yields of diverse crops. The efficacy of deep learning network frameworks has been underscored by deploying several ML models for estimating agricultural yields. Common examples of these models include SVM, RF, and LASSO [10].

Crop output is subject to the influence of climate, management strategies, and genotypic factors. Large-scale meteorological events can significantly impact crop growth and yields by altering regional climatic patterns, as highlighted in studies such as [11] and [12]. Given the pivotal role of environmental conditions in crop development across various stages, utilizing a diverse set of environmental parameters for predicting crop yield becomes crucial, as emphasized by [13]. Furthermore, satellite imagery can detect biotic variables, including diseases and insects, which can impact crop development and manifest in leaf traits, as discussed by [14]. Incorporating remote sensing measures that enable real-time crop growth status monitoring can enhance yield forecasts' accuracy.

Crop production predictions at regional, national, and global scales have commonly employed meteorological data, remote sensing data, or a combination of both, as evidenced by studies such as [15], [16], and [17]. Prediction models typically integrate primary meteorological and satellite-based input variables, including temperature, precipitation, solar radiation, and the Normalized Difference Vegetation Index (NDVI), as outlined in works [18], [19]. Despite its importance as a climatic factor affecting plant growth through its influence on foliar gas and heat exchange, modification of foliar boundary layers, and alteration of water status, wind speed has received comparatively less attention in these prediction models [20], [21].

Process-based models demand extensive data inputs and calibration criteria, as emphasized by [22]. As an illustration, the development of satellite-based light use efficiency models, assuming that gross primary production is solely determined by the amount of photosynthetically active radiation, aimed to estimate vegetation gross primary production [23]. In a comparative study assessing statistical performances, [16] discovered that Random Forest outperformed multiple linear regression in all evaluated metrics, showcasing its potential for predicting agricultural productivity [16]. Although ML has demonstrated success in numerous large-scale agricultural yield prediction studies in China, as highlighted by [24], its suitability for national-scale wheat yield prediction remains unexplored [25].

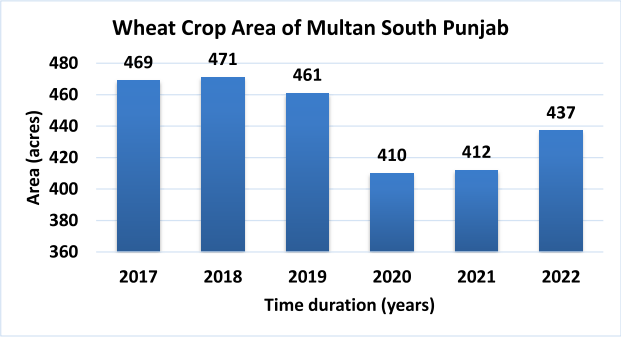

FIGURE 1. Multan wheat crop area from 2017 to 2022.

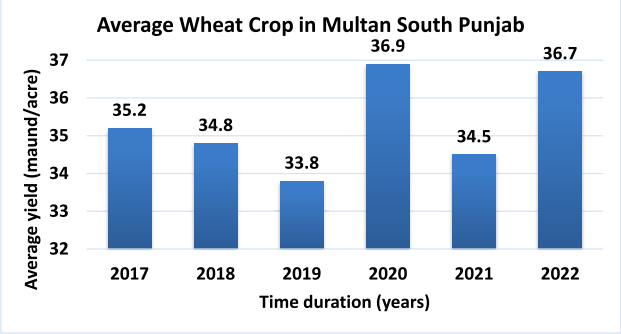

FIGURE 2. Multan wheat crop area from 2017 to 2022.

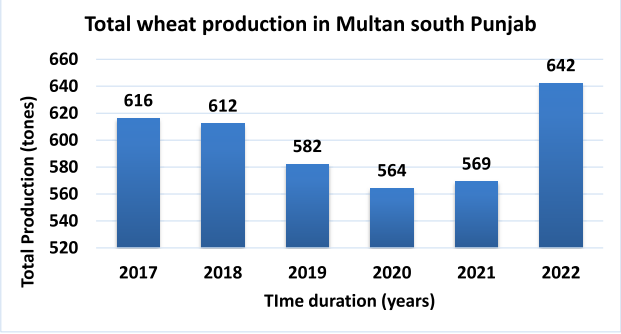

FIGURE 3. Wheat crop production in Ton from 2017 to 2022.

This study compared multitemporal Moderate Resolution Imaging Spectroradiometer (MODIS) enhanced vegetation index (EVI) and NDVI data to estimate rice crop yields in the Mekong River Delta (MRD) of Vietnam [26]. In this study, the author tried decision trees and Random Forests for crop yield prediction and a Decision Support System for Agro-Technology Transfer (DSSAT) simulation for crop yield prediction [27]. The primary goal of this study was to

assess the applicability of monthly composites from Sentinel-2 images for rice yield prediction at the field size in Taiwan using ML approaches [28]. This work used Landsat 8 surface reflectance products from 2017, 2018, and 2019 to map the satellite-based NDVI, leaf area index (LAI), and normalized difference water index (NDWI) [29].

This study aims to create an ML approach for predicting rice crop yields in Taiwan using time-series Moderate Resolution Imaging Spectroradiometer (MODIS) data [30].

The main aims are (a) to evaluate the prediction accuracy of the four ML algorithms, ANN, SVR, KNN, and RF, and (b) to evaluate the impact of different feature sets on ML algorithms. The following feature selection algorithms are used to identify unique feature sets: Forward Feature Selection (FFS), Correlation-based Feature Selection (CBFS), Variance Inflation Factor (VIF), and Random Forest Variable Importance (RFVarImp) [31].

To create a county-level corn yield prediction model based on Bayesian Neural Network (BNN) employing numerous publically available data sources, such as time-series satellite products, sequential climatic measurements, soil property maps, and historical maize yield records [32].

Remote sensing (RS) systems are increasingly used to develop decision support tools for modern farming systems, aiming to improve yield output and nitrogen control while lowering operating costs and environmental effects. However, RS-based systems require massive amounts of remotely sensed data from many platforms. Therefore, more attention is increasingly being paid to ML methods. This is owing to the ML system's ability to process many inputs and handle nonlinear tasks [33]. As a result, the scholarly literature was screened to identify a wide range of essential features for capturing current progress and trends, such as (a) the research areas most interested in ML techniques (RF, SVM, LASSO) in agriculture, as well as the geographical distribution of the contributing organizations, (b) the most efficient ML models, (c) the most investigated crops and animals, and (d) the most implemented features and technologies [34].

Deep learning techniques are typically unsuitable for general-purpose applications since they require vast data. Tree ensembles typically outperform them for traditional ML issues. Furthermore, they are computationally intensive to train and need significantly more expertise to tune (i.e., setting the architecture and hyper-parameters) [35].

Conventional or outdated approaches to yield estimation are both labor-intensive and time-consuming. Additionally, the collection of yield data from a limited number of villages fails to adequately represent the broader agricultural landscape [3]. Although past research has effectively utilized various models for predicting wheat and other crop yields, there is a clear emphasis on carefully selecting base learners and predictors to enhance model performance. For instance, a study employed a step-wise multiple regression method to develop models for predicting departmental-level wheat yield in France [36]. Predicting wheat crop yield in Pakistan poses challenges due to its dependence on a multitude of internal factors (such as seed, disease, and plant health), external factors (including weather, soil, irrigation, and socio-economics), and the reliance on manual methodologies, leading to imprecise estimates before harvesting.

This study aimed to investigate meteorological constraints on winter wheat output and develop a yield forecast model based on meteorological data [37]. This study proposes an approach using three ML models for predicting crop yield. This technique utilizes remote sensing images and various factors as input, and the input data undergoes validation in a software engineering process [8]. The methodology presented in this study holds significant value for policy-makers in improving the identification of import strategies and making decisions to address large-scale food security challenges in Pakistan [38].

The general contributions of the article are as follows.

- Exploring the viability of ML-based yield forecasting using historical yield data.
- Showcasing the significance of yield detrending and the influence of climate, soil, and socioeconomic factors on yield estimation
- Minimize Pakistan's local conventional manual wheat crop yield estimation process.
- This study would make possible accurate crop yield prediction of wheat before the harvesting season.

The remainder of the paper is structured as follows: Section II covers the Materials and Methods employed in the study, while Section III delves into the proposed approach. Section II-C presents results and analysis of the proposed approach, while section V concludes the article.

II. MATERIALS AND METHODS

This section describes the meteorological characteristics, Datasets, and data sources.

A. STUDY REGION

Data on wheat were gathered for the inquiry in the Punjab province of Pakistan from 2017 to 2022. The dates of planting, growth, overwintering, returning green, jointing, smooth development, and reaping were included in this information. The daily leaf territory list and soil field capacity were computed using these dates, and the water-restricted potential generation was calculated using those results. A physically based HYDRUS-1D and linear regression models were used to analyze the correlation between meteorological parameters and temperature at different depths in silt loam soil [39]. Here are some weather statistics for the wheat-growing season. High-density raster photographs of the target region, located at latitude 29.848212 N and longitude 71.263367 at 423 feet (129 m) above sea level, were among the data made available for this study. According to Crop Report, it has a semi-arid climate typical of Multan District, Punjab, Pakistan, with an area of 3,721 square kilometers and a total of 437 acres according to Crop Report Services Punjab shown in Fig. 4. The terrain is flat and alluvial, making it suitable for agriculture. The irrigation network of canals makes the

area extremely fruitful. During the monsoon season, land adjacent to the Chenab River is typically inundated. All areas with sweltering summers and moderate winters have an arid climate. The average rainfall is approximately 186 mm (7.3 in). The location sees some of the most extreme weather in the country, with a maximum recorded temperature of around $52^{\circ}C$ $(126^{\circ}F)$ during 2010 and the lowest recorded temperature of roughly $-1^{\circ}C$ $(30^{\circ}F)$ [38]. Dust storms are a typical occurrence in this area. Multan is located in the cotton-wheat farming zone of Punjab. Wheat is the principal crop of the Rabi season, and cotton is the major crop [40].

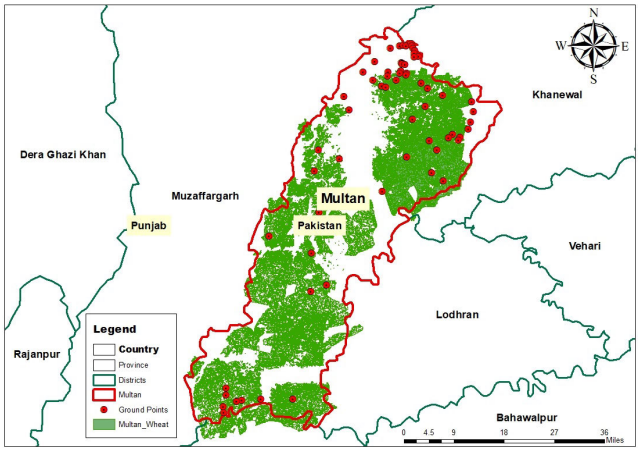

FIGURE 4. Study Region

B. DATA COLLECTION

In this study, the wheat crop yield production dataset has been realistically obtained, encompassing details such as wheat crop area, wheat crop production per acre, and overall wheat production in Ton for the Multan district over the last six years. This data was sourced from the Crop Reporting Services of the Government of Punjab, Pakistan. The inquiry focused on wheat data in the Punjab province of Pakistan, covering 2017 to 2022.

Fig. 1 illustrates the wheat crop area in Multan from 2017 to 2022, provided by Crop Report Services Punjab, Pakistan. Fig. 2 shows the wheat crop yield per acre for the same period and location, based on Crop Report Services Punjab data. Additionally, Fig. 3 displays the total production of wheat crops in Multan from 2017 to 2022, sourced from Crop Report Services Punjab, Pakistan.

The data utilized in this study was acquired through the POWER Data Access Viewer, accessible at https://power.larc.nasa.gov/data-access-viewer/ the end of March for each year from 2017 to 2022. In the cross-validation process, meteorological data was additionally downloaded from the Pakistan Meteorological Department of the Pakistani government. This data is particularly crucial for every agricultural output, with staple crops such as wheat, maize, cotton, rice, and sugar cane (listed in Table 1) facing significant risks due to climate change [41].

The anticipated rise in temperatures is a major concern. By the year 2040, a temperature increase of 3°C is projected, and by the end of the century, this rise is expected to reach $5-6$°C. Such temperature changes could potentially lead to a 50% reduction in wheat output in Asian nations. Given its geographical location, the impact on Pakistan is predicted to be even more severe. The features considered for this research encompass the various elements that contribute to climate change, posing a threat to the crop production systems of staple crops such as wheat, corn, cotton, rice, and sugar cane.

The NDVI [42], obtained from the United States Geological Survey (USGS), is a simple graphical indicator that can be used to analyze remote sensing measurements and determine whether the target being observed contains live green vegetation as shown in Eq. (1).

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

The NDVI is determined as a ratio of the total and difference of two Landsat bands: near-infrared (NIR) and red. The NDVI is the most well-known and commonly used ecological indicator for measuring vegetation conditions, thanks to an effectively designed formula incorporating the spectral features of red and NIR bands. Thus, compared to the other bands, the NIR band strongly reflects the chlorophyll content of healthy vegetation. At the same time, it absorbs red light. Using these two bands in a well-established combination yields robust data revealing the presence and amount of chlorophyll in leaves [43]. Typically, but only sometimes, this analysis is done from a space platform. The selected NDVIs are used to develop yield forecasting and the overall process for classification of all images shown in Fig. 5.

C. EXPERIMENTAL SETUP

Initially, Google Earth Engine (GEE) is employed to implement SVM, Random Forest, and LASSO, both powerful ML algorithms. These models are efficiently applied to analyze large-scale datasets, leveraging GEE's cloud-based processing capabilities, making it an ideal platform for handling remote sensing data. The SVM and Random Forest models are trained to predict wheat yield, utilizing a combination of spectral indices and spatial attributes. Subsequently, Lasso Regression is employed in ArcMap for feature selection and regression. The analysis results in thematic maps illustrating the spatial distribution of predicted wheat yields across the study area. Regression graphs are generated using Excel to visualize the performance of Lasso Regression in predicting yield values, comparing them with actual ground truth data [44]. Fig. 7 shows the Land Use and Land Cover LULC of wheat crops in the Multan district from 2017-2022, and Fig. 10 compares wheat crops and other things in Multan that are extracted from SVM. All models utilize data inputs, including remote sensing data collected throughout the growth season and climatic data collected

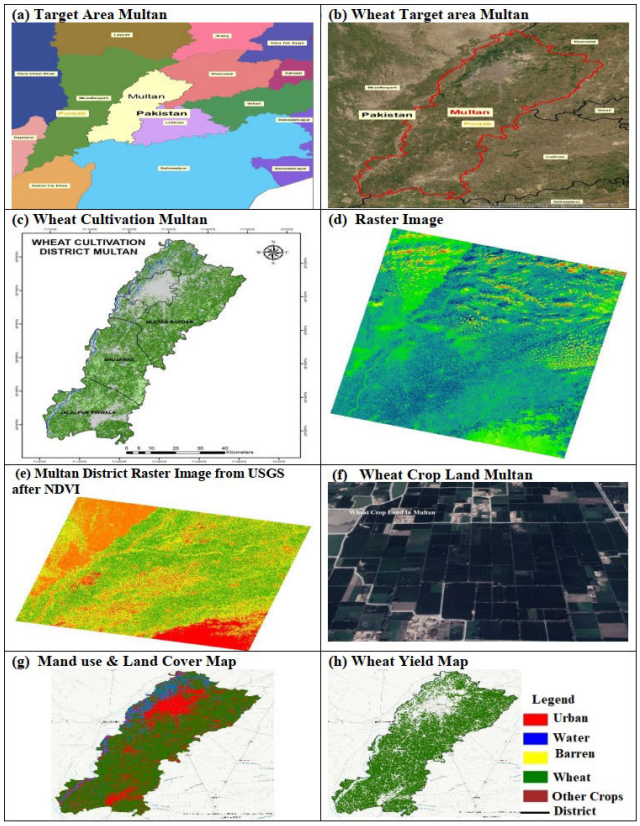

FIGURE 5. Study area map of different remote sensing stages, a shows Multan district in Punjab, b shows Multan google map area, c shows Multan district wheat crop area, d, e, f, g, and h shows the different classification of NDVI map of the wheat area from 2017 to 2022.

during a specified period. In this instance, data entered during a particular time or the complete growing season refers to monthly data throughout the growing period, as opposed to the average value. The predicted performance is compared to the benchmark provided by CRS Punjab, Pakistan, which employs remote sensing data collected throughout the month.

III. PROPOSED APPROACH

For this investigation, location data for the understudied location (423 feet (129 m) above sea level) were available. According to Crop Report Services Punjab, it is located in the semiarid Multan District of Punjab, Pakistan, and has 437 acres and a land area of 3,721 square kilometers. Every year, there is an average of 304 millimeters of rain and 26.2 degrees Celsius. In terms of precipitation, summers are wetter than winters. The framework flow diagram Fig. 6 decreases in November and increases in January. Multan's top layer of soil is made up of fine sand, known as silt, as shown in Fig. 5. Multan's mixed-cutting zone grows sugarcane, maize, wheat, and rice. Fig. 6 shows the overall framework process in a scientific model that shows a Raster Landsat image is collected from satellite USGS, then processed that raster image is to calculate NDVI values for the dataset, after that collet factor dataset from different sources, and then ML techniques apply that dataset after Training the dataset with ML techniques and finding results in the form of predicting wheat crop yield. It tends to rain more heavily in the summer than in the winter.

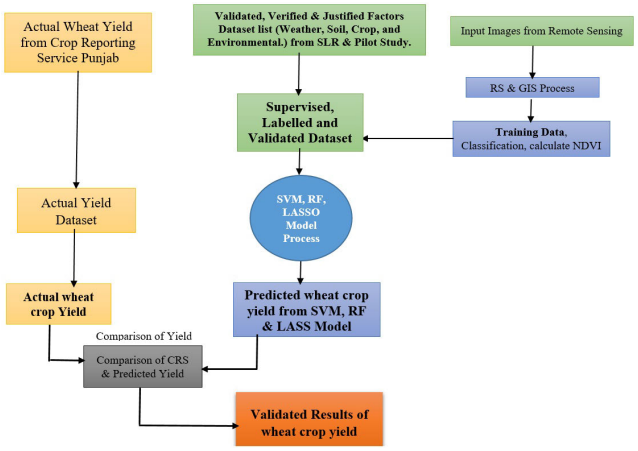

FIGURE 6. Framework flow diagram.

A. DATA PREPROCESSING

The processes involved in data processing and yield forecasting encompass various stages, starting with data collection and progressing through processing, calibration, and validations. A systematic approach was followed in conducting these processes, commencing with data collection and preprocessing. The sequence then advanced through vegetation indexing and smoothing, NDVI data masking based on fields, calibration of the masked data with yield information, and ultimately, validation of the yield forecasting models.

The NDVI data in this study was motivated by their significant correlations with wheat yield, as demonstrated in previous research. A smoothing technique based on penalized splines was applied to mitigate noise in the NDVI time series data. The NDVI values were derived using Eq. (1) from Landsat 8 satellite images. Google Earth Engine (GEE), a planetary-scale platform for earth science data and research, facilitated remote sensing data management.

Further processing was conducted on meteorological and soil data acquired and analyzed on a desktop computer. An 8-day interval time series was generated using a maximum-value composite to minimize noise, and simultaneous cropland masking was applied. Meteorological data were aggregated every eight days to align with the period of the remote sensing data. Spatial aggregation to the mean for each district was performed for all data, encompassing remote sensing, meteorological, and soil data, using district boundaries. Invalid data points were eliminated and interpolated, and each variable's maximum and minimum values were standardized to a range of 0-1. Consequently, the resulting time series had a consistent length of 32 each year, covering the entire winter wheat growing season from November

to April of the following year. By combining satellite data, remote sensing data, and satellite imaging, researchers can gain a comprehensive and timely understanding of the environmental conditions affecting wheat fields. This strategy encourages farmers and agricultural stakeholders to make more informed decisions [42]. From 2017 to 2022, researchers gathered information on the region's weather and climate as well as its soil and crop yields and distribution.

Table1 presents a summary of all the input variables. The study started by resampling the data for a spatial resolution of districts and a temporal resolution of months. Then, more cover was added based on where the winter wheat was planted. Ultimately, the study arrived at district-level averages for all of the Multan. ArcGIS and the GEE platform were used to bulk up the analysis.

TABLE 1. All input values (factors) and source.

Category	Variables	Duration	Source
Crop,Wheat Area	Yield,area	Yearly	CRS
Satellite Data	NDVI	Yearly	USGS
Climate Data	Weather Data	Daily	Metrological Dept
Soil Data	soil Type, pH	Yearly	Soil Fertility Authority

B. MACHINE-LEARNING METHODS FOR PREDICTING WHEAT CROP YIELD

For ML, this study utilized the Anaconda Jupyter Notebook with Python (version 3.6.2) and statistical software [45], implementing a framework facilitated by the caret package [46]. This package offers a consistent interface for various models developed using other Python packages. The ML techniques employed included LASSO, RF, and SVM, all nonlinear methods.

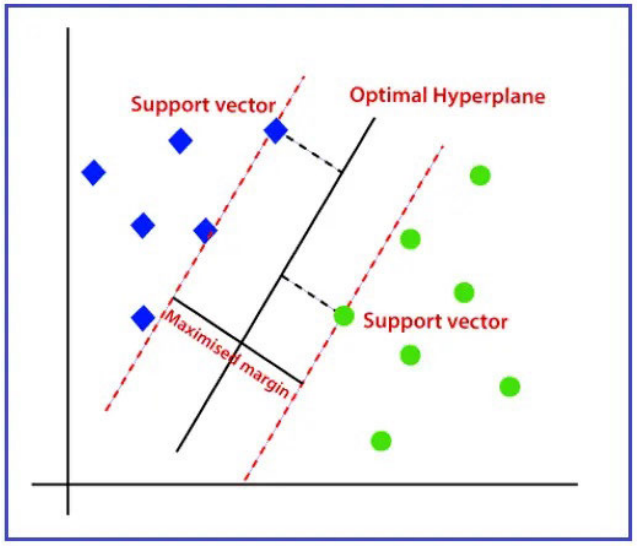

FIGURE 7. SVM algorithms [37].

1) SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised, non-parametric approach that uses marginal kernels [45]. When minimizing errors without overfitting the model, SVM regression uses a kernel function to map the input onto a high-dimensional feature space, where a linear regression model is subsequently built. Among the most crucial hyper-parameters to tweak are the kernel functions used in the model. You can apply the kernel trick to capture nonlinear relationships between input features. The decision function with a kernelized SVM becomes:

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x) + b \tag{2}$$

Here, α_i are the Lagrange multipliers obtained during optimization, and $K(x_i, x)$ is the kernel function in Eq. (2).

In your specific case, you'd need to define the features (xx) based on the data you have shown in Fig. 7. These features could include rainfall, temperature, soil quality, and other relevant agricultural parameters for wheat crop yield prediction.

This is a general representation; the implementation would depend on the dataset and specific requirements. Consult with agriculture or data science experts to fine-tune the model based on domain-specific knowledge and data availability.

2) RANDOM FOREST (RF)

Random Forest (RF) is a nonparametric method for regression tree analysis and advanced categorization, known for its resilience against overfitting and effectiveness with high-dimensional datasets [45], [47]. On the other hand, SVM is a supervised learning model utilized for regression and classification tasks [47]. This study adopted a high-dimensional feature space, employing a kernel function (linear, Gaussian, polynomial, or hyperbolic tangent) for SVM regression. Specifically, the Gaussian kernel function was selected to explore the nonlinear relationship between input predictors (climate and remote sensing data) and output predictors (yield) [48]. Random Forest enhances predictive accuracy and mitigates overfitting by amalgamating multiple decision trees. The ensemble nature involves training numerous trees on distinct subsets of the data and averaging their predictions. Let's denote the Random Forest model as F(x), where x represents the input features for wheat crop yield prediction.

Let's denote the Random Forest model as F(x), where x represents the wheat crop yield prediction input features.

$$F(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{3}$$

Here, N is the number of trees in the forest, and $f_i(x)$ is the prediction of the i^{th} decision tree in Eq. (3).

The prediction of an individual decision tree $f_i(x)$ is obtained by traversing the tree based on the input features. Each leaf node in the tree represents a predicted value. Scikit-learn rest measures feature importance based on how much

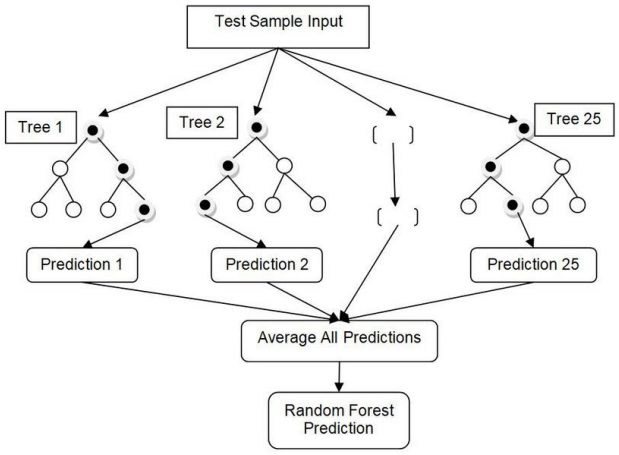

FIGURE 8. Random Forest (RF) classifier process diagram.

each feature contributes to the reduction in impurity (e.g., Gini impurity) across all trees, as shown in Fig. 8.

Random Forest has hyperparameters that need to be tuned, such as the number of trees (NN), the maximum depth of each tree, and the number of features considered at each split. In practice, the implementation involves training the Random Forest on historical data with known crop yields and then using the trained model to predict the yield for new data. This mathematical representation provides an overview of the Random Forest model for wheat crop yield prediction [22]. The actual implementation would involve using an ML library (e.g., sci-kit-learn in Python) and adjusting parameters based on the characteristics of your dataset.

3) LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO)

This study's LASSO technique is a shrinkage and selection approach for linear regression. It minimizes the residual sum of squares, allowing the total of absolute coefficient values to be less than a defined value [41]. Because of its automatic feature selection, LASSO is particularly well suited for parsimonious regression models, alleviating the problem of overfitting input data. LASSO is a linear regression regularization approach that promotes sparsity in model coefficients while preventing overfitting. It entails incorporating a penalty element into the linear regression goal function.

The decision function for LASSO is the linear combination of the input features and their corresponding coefficients:

$$f(x) = wx + b \tag{4}$$

The features (xx) could include various factors relevant to wheat crop yield prediction, such as weather conditions, soil quality, and agricultural practices in Eq. (4).

The goal is to find the values of w and b that minimize the objective function. The optimization problem involves balancing the fit to the training data (first term) with the regularization term to prevent overfitting and encourage sparsity.

LASSO is especially beneficial for dealing with high-dimensional datasets requiring feature selection for several features. The L1 regularization term supports a sparse solution, which might be advantageous for selecting only the most relevant ones for prediction. Before further analysis, all predictors and yield were normalized with a mean of zero and a standard deviation of one. The dataset could calculate predicted values for R^2 and RMSE in optimal models.

IV. RESULTS AND DISCUSSIONS

This section presents the experimental setup and results obtained using the three ML models for wheat crop yield prediction in the focused region.

To analyze the distinctive contribution of remote sensing data at a specific timeframe, inputs from various times for both all-month climate data and remote sensing data are included. The distinction between using climate and remote sensing data together during the growth season and using climate data alone or remote sensing data exclusively is underlined.

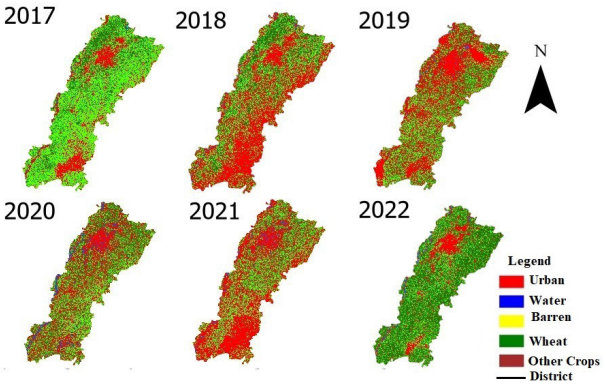

FIGURE 9. Show land use and land cover LULC of wheat crop in Multan district from 2017-2022.

A "leave-one-year-out" forecast is then utilized to assess how well the approaches perform outside the sample, providing insights into the models' generalizability.

A. SVM FOR WHEAT CROP YIELD ESTIMATION

SVM is a robust supervised ML algorithm extensively employed for tasks such as classification and regression. In wheat yield estimation, SVM is a valuable tool for forecasting crop yields by leveraging input features and historical yield data. The underlying principle of SVM involves identifying the hyperplane that optimally separates data points corresponding to distinct classes. In the context of regression, SVM seeks to find the hyperplane that maximizes the difference between predicted and actual values. SVM is a strong supervised ML method commonly used for classification and regression tasks. SVM is a powerful technique for forecasting crop yields based on input parameters and historical yield data in wheat yield estimate [48].

SVM's main idea is to locate the hyperplane that best divides data points from distinct classes. In regression, SVM seeks the hyperplane with the most significant margin between the predicted and actual values. The kernel functions in the model depicted in Fig. 9 are among the most important hyper-parameters to fine-tune.

1) TRAINING DATA GENERATION

The training dataset has been generated by merging feature collections representing various land classes such as water, urban areas, barren land, wheat fields, and other crops. Relevant bands from Landsat 8 imagery, including B2 (blue), B3 (green), B4 (red), B5 (near-infrared), and B7 (shortwave infrared), were extracted. This training dataset encompasses spatial attributes and corresponding land-use land-cover (LULC) labels. Employing the SVM model on Landsat 8 imagery enabled wheat yield prediction across the study area shown in Fig. 10, [18]. The resulting classified map visually presents the spatial distribution of wheat yield values, offering insights into variations in crop productivity. The outcomes are visually represented using a color palette to signify different wheat yield levels and other crops shown in Fig. 11.

Over the six years, from 2017 to 2022, wheat cultivation in Multan District displayed fluctuating trends concerning area and yield. Despite these variations, the wheat yield per acre exhibited a more stable pattern, ranging from 35 maunds per acre in 2019 to 41 maunds per acre in 2018. This suggests that while the cultivated area changed over the years, the productivity per acre remained relatively consistent. The total yield, accounting for both area and per acre yield, underwent similar fluctuations. The total yield peaked in 2018 at 932.01 Tons and reached its lowest point in 2022 at 639.80 Tons. Notably, the total yield in 2022 decreased despite a relatively high per acre yield of 37 maund per acre. This indicates that the reduced cultivated area significantly impacted overall wheat production and showed R^2 in Fig.12. Table 2 provides basic information about the dataset utilized in this study.

TABLE 2. Wheat yield in Multan district (2017 - 202 predicted by SVM.

Year	Area (acre)	Per acre Yield(Maund)	Total Yield (Ton)
2017	543.7	39	848.17
2018	568.3	41	932.01
2019	519.7	35	727.58
2020	519.7	36	748.37
2021	531	38	807.12
2022	432.3	37	639.80

The SVM-based wheat yield estimation demonstrated promising results, enabling accurate predictions based on spatial attributes. The classified map highlights regions with high and low wheat productivity, aiding agricultural decision-making. The successful application of SVM on GEE for wheat yield estimation opens avenues for further research in ML and remote sensing applications in agriculture.

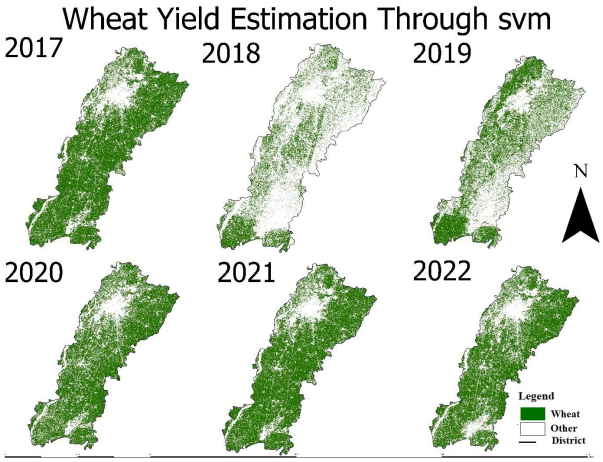

FIGURE 10. Visual representation of wheat crop yield prediction of wheat using SVM from 2017 to 2022 map.

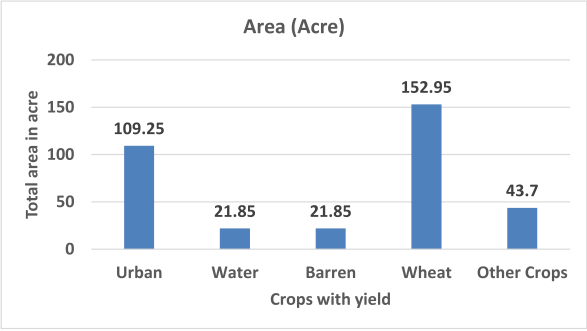

FIGURE 11. Show a comparison of wheat crop area (acre) and other things in Multan that extract from SVM.

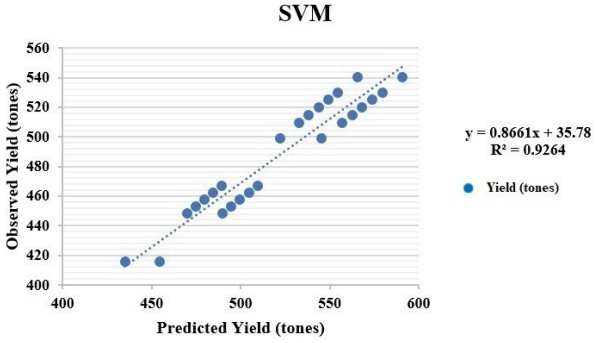

FIGURE 12. Wheat crop yield pattern of target area Multan district from 2017-2022 using SVM results.

B. RANDOM FOREST (RF) FOR WHEAT CROP YIELD ESTIMATION

Random Forest is an ensemble learning technique that builds multiple decision trees during training and amalgamates their predictions to generate more robust and accurate results. This method addresses several challenges inherent in individual decision trees, such as overfitting and bias, by averaging the predictions of numerous trees. Given the diverse factors influencing crop productivity, Random Forest is a valuable

choice for wheat yield estimation. It is particularly effective for intricate datasets with high-dimensional feature spaces and a substantial number of training samples.

This study acquired USGS Landsat 8 Collection 2 Tier 1 TOA Reflectance imagery through Google Earth Engine (GEE), covering the expanse of Multan District. To ensure the quality of the data, filtered for cloud-free scenes spanning from March 2017 to 2022 for time series analysis. The region of interest (ROI) corresponds to Multan District, and the spatial attributes of wheat fields were derived from the Landsat 8 imagery. The relative importance of measured variables may be quantified with RF, and it is an efficient method for selecting relevant variables [18] shoe in Fig. 13.

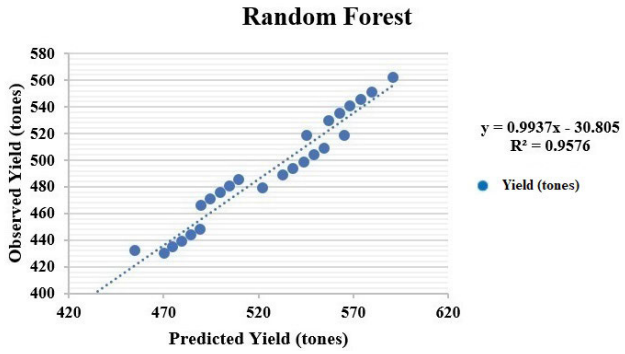

FIGURE 13. Wheat crop yield pattern of target area Multan district from 2017-2022 using RF results.

1) TRAINING DATA GENERATION

A training dataset combines feature collections representing water, urban areas, barren land, wheat fields, and other crops. Extracting pertinent bands from Landsat 8 imagery, such as B2 (blue), B3 (green), B4 (red), B5 (near-infrared), and B7 (shortwave infrared) as illustrated in Fig. 14, establish the training dataset with spatial attributes and associated Land Use and Land Cover (LULC) labels.

2) WHEAT YIELD ESTIMATION AND VISUALIZATION

The trained Random Forest model is applied to the Landsat 8 imagery to predict wheat yield across the Multan District. The resulting classified map visually represents the spatial distribution of predicted wheat yield values. The study employs a color palette to distinguish different levels of wheat productivity. This study compiles wheat production data for 2017 to 2022 to assess wheat yield trends. The table presents the wheat area (in acres), yield per acre (Ton), and total yield (Ton) for each year shown in Fig. 12 and Table 3. Analyze the temporal variations and growth in wheat yield to gain insights into the factors contributing to increased productivity [49].

In this table, area in areas per area yield in maund and Total Yield in Tone.

The total yield, which accounts for the area and per acre yield, experienced similar fluctuations. It reached its highest point in 2018, at 21.02 Ton, and dipped to its lowest in 2022, at 496.08 Ton. Notably, the total yield in 2022 decreased even though the per acre yield was relatively high at 36.08 and most deficient at 31.68 mounds per acre.

TABLE 3. Wheat crop yield data (2017-2022) predicted by RF.

Year	Area(acre)	Per acre Yield(Maund)	Total Yield (Ton)
2017	479.6	34.32	658.39
2018	499.6	36.08	721.02
2019	457.6	30.8	563.76
2020	457.6	31.68	579.87
2021	466.9	33.44	624.53
2022	380.9	32.56	496.08

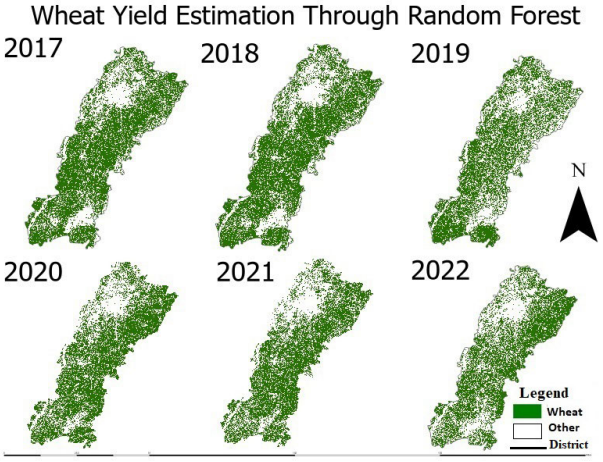

FIGURE 14. Show visualization of wheat crop yield pattern of target area Multan district from 2017-2022 using RF.

C. LASSO REGRESSION FOR WHEAT CROP YIELD ESTIMATION

Lasso Regression is a linear regression technique that incorporates L1 regularization. In wheat crop yield estimation, Lasso Regression is employed to model the relationship between input features and the yield output while performing feature selection. The L1 regularization term in Lasso Regression encourages sparsity in the coefficient values, effectively setting some coefficients to zero.

The key objective of Lasso Regression for wheat crop yield estimation is to find the optimal set of coefficients that minimizes the sum of squared differences between the predicted and actual yield values while penalizing the absolute values of the coefficients. This helps identify the most influential features contributing to the yield prediction. In wheat crop yield estimation, Lasso Regression aids in identifying and prioritizing the significant factors among various inputs, such as climate data, soil attributes, and management practices, that impact the overall crop productivity [41]. This technique is beneficial when dealing with high-dimensional datasets, where feature selection is crucial for model interpretability and efficiency, as shown in Fig. 15.

Overall, Lasso Regression is a valuable tool for wheat crop yield estimation by providing a balance between accurate prediction and feature selection, helping identify the most relevant factors influencing crop yield.

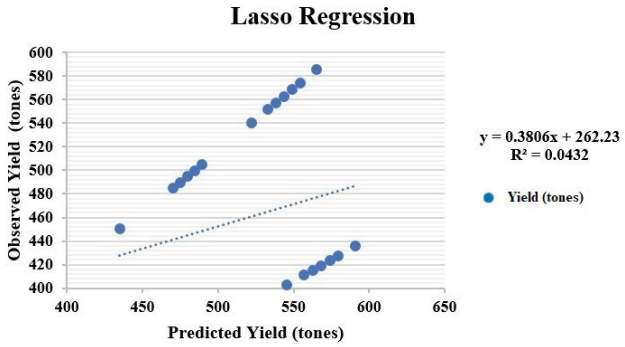

FIGURE 15. Wheat crop yield pattern of target area Multan district from 2017-2022 using LASSO results.

TABLE 4. Wheat crop yield data (2017-2022) using LASSO.

Year	Area (acre)	Per acre Yield(Maund)	Total Yield (Ton)
2017	445.9	31.91	569.15
2018	465.2	33.51	623.55
2019	425.3	28.63	487.05
2020	425.3	29.4	500.15
2021	435.1	31.08	540.92
2022	353.8	30.25	428.10

Analyzing the time series data from 2017 to 2022 provides insights into wheat production trends and growth patterns. Table 4 details wheat area (in acres), yield per acre (Ton), and total yield (Ton) for each year, offering a comprehensive view of changes in wheat productivity over this period.

The total yield exhibits fluctuations, accounting for the cultivated area and per-acre yield. The peak total yield occurred in 2018, reaching 623.55 TTons, while the lowest point was observed in 2022, with a total yield of 428.10 TTons. An interesting observation is that the total yield in 2022 experienced a decline despite a relatively high per-acre yield ranging from 31.68 to 36.08 maund per acre. This suggests that the reduction in the cultivated area significantly impacted the overall wheat production during that year. A time series of wheat yield data for 2017 to 2022 is analyzed to examine yield trends and growth in wheat production. Table 4 presents the wheat area (in acres), yield per acre (Ton), and total yield (Ton) for each year, allowing us to observe the changes in wheat productivity over time.

Total yield, which accounts for the area and per acre yield, experienced similar fluctuations. The total yield reached its highest point in 2018, at 623.55 Ton, and dipped to its lowest in 2022, with 428.10 Ton. Notably, the total yield in 2022 decreased even though per acre was relatively high at 36.08 and most deficient at 31.68 maund. This indicates that the reduction in the cultivated area significantly impacted overall wheat production, shown in Table 4.

D. METRICS FOR MODEL EVALUATION

When deciding between the three ML and DL methods, cross-validation (CV) is a popular tool for algorithm selection due to its ease of use, general applicability, and ability to prevent over-fitting [1]. It is common practice to favor the model with the smallest relative estimation error. Five-sample cross-validation was used to pick the best models for this analysis. This study compared the ML model's accuracy using several metrics, including the RMSE in Eq. (6), the coefficient of determination (R^2) in Eq. (5), and the mean absolute error (MAE) in Eq. (7).

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y_i})(f_i - \bar{f_i})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2 \sum_{i=1}^{n}(f_i - \bar{f_i})^2} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2} \tag{6}$$

$$MAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - f_i|} \tag{7}$$

The variables y_i, $\bar{y_i}$, f_i, and $\bar{f_i}$ indicate the observed and predicted winter wheat yields, whereas n denotes the number of samples used in the ML model ($i = 1, 2, \cdots, n$). Predictive ability increases as R^2 in Eq. (2) approaches 1.

Less of a gap exists between actual and predicted yields when the RMSE in Eq. 3 and MAE in Eq. 4 have smaller values. The study computed the root-mean-squared error, correlation, and mean absolute error to evaluate the accuracy of the model's predictions. Research [10] demonstrates that the Global Moran's I measure for assessing the spatial autocorrelation of prediction errors supports model generalizability throughout the geographical domain. In the range of -1 to 1, the global Moran's I value can indicate a tendency towards clustering or spreading out, respectively. Spatial randomness is displayed by values close to zero, and more uncertainty indicates a more robust model [50]. In this research, In this study used ArcGIS 10.3 and GEE to estimate the worldwide value of Moran's I [50].

E. PERFORMANCE EVALUATION

Compared to the previous study that integrated climate and remote sensing data for winter wheat yield prediction in Multan, the latest forecasted R^2 values (ranging from 0.74 to 0.88) in the district (Fig. 10 showed an improvement. Notably, temperature data exhibited the highest yield prediction accuracy, aligning with findings from prior research. The relatively lower yield prediction performance of remote sensing data compared to climate data might be attributed to

mixed signals from land surfaces, particularly during sowing and seedling emergence, in contrast to climate data, which directly corresponds to the overall climatic conditions during the wheat growth season. Considering the provided RMSE values and their ranking, Random Forest and SVM emerged as the most effective methods for wheat yield estimation. At the same time, Lasso Regression was deemed less optimal for this particular prediction task.

TABLE 5. Show a comparison of the wheat crop in tone with benchmark CRS (Benchmark) Punjab Pakistan.

Year	SVM(tones)	RF(tones)	LASSO(tones)	CRS(tones)
2017	848.172	658.3949	569.1468	616
2018	932.012	721.0227	623.5541	612
2019	727.58	563.7632	487.0536	582
2020	748.368	579.8707	500.1528	564
2021	807.12	624.5254	540.9163	569
2022	639.804	496.0842	428.098	642

After all results and comparisons, it was concluded that SVM and RF produced the best result in this study area within stud area factors like temperature, precipitation, humidity, soil, etc. Table 5 compares all techniques with CRS data of total production in the whole district of Multan from 2017 to 2022.Fig.16 shows the accuracy in percentage in the scenario of the study area. With current factors, the result shows RF is the best classifier that returns 97% accuracy, and SVM shows 93%, which is also good accuracy. Results show that LASSO has the lowest accuracy, 85%, because of each classifier specification in the study area.A comparison of the SVM, RF, and LASSO calibrated models indicated significant differences in their performance when forecasting wheat production using Landsat-8 data, soil information, and topography features from other regions using z-test satanical method [51]. Given the encouraging and diverse results achieved of wheat yield estimation, the proposed methodology has the potential for implementation in other study regions with similar agricultural systems and climatic conditions.

TABLE 6. Comparison of ANN predicted yield and CRS observed yield.

Models	R^2	MAE	RMSE
Random Forest	0.8825	0.5800	0.05617
SVM	0.7450	0.8217	0.0317
Lasso Regression	0.8000	0.8183	0.6025
Average	0.8091	0.740	0.2301

The correlation analysis indicated potential relationships between the yield and various environmental factors, particularly from late February to March. This time frame aligns with the grain-filling phase before the harvest in many regions of Multan. Notably, there was a negative correlation between rainfall and yield in extensive areas. Intense and prolonged rainfall events during this period can occasionally lead to pre-harvest sprouting, negatively impacting grain quality and reducing the salable yield. This contrasts with other wheat-producing regions globally, where drought is more commonly associated with yield reduction.

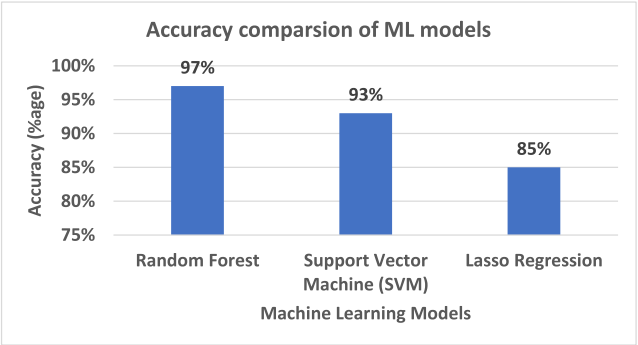

FIGURE 16. The overall accuracy of three ML techniques.

TABLE 7. Comparison with existing studies.

Existing study	Factors	Accuracy
Proposed Framework	Soil, Water, NDVI, Weather	97%
[5]	Genotype, Environment	91%
[11]	Field initial data and Management	93%
[14]	Climate Factor, genotypes	90%
[22]	genotypes	80%
[44]	Weather conditions, Remote Sensing	95%
[37]	Crop Management factors	87%

F. COMPARATIVE ANALYSIS

Its accuracy must be assessed to establish how well a suggested wheat crop production prediction framework compares to alternative ML and DL approaches. Table 7 corresponds the proposed strategy to the state of the art to demonstrate the generalizability of the approach. Accuracy is frequently the most critical factor when evaluating ML systems. Other performance indicators, such as accuracy, were also used to assess the classifier's performance under various conditions. The suggested deep and ML model significantly outperforms state-of-the-art approaches.

Machine learning algorithms like ANN, LSTM, and RNN have recently gained popularity for time-series forecasting. The ANN model was identified as a crucial method. ANNs discover patterns in data and generate generalizable solutions based on existing information. Research indicates that ANNs are most commonly used in the financial sector. Recent research in time-series forecasting has focused on using RNN and its derivatives, including LSTM [52]. Recent advancements integrate DL and ML algorithms to generate complex data. This study aimed to compare the performance of machine learning approaches with non-linear algorithms, identifying their advantages and weaknesses. Using it reduces computing time and ensures high-quality work [53].

Both the recommended technique and the cutting-edge procedures are somewhat inaccurate. The proposed method

was evaluated using wheat prediction methodologies currently in use in the literature. As demonstrated in Table 7, the proposed technique achieved the highest accuracy (97%) compared to recent investigations. The suggested framework surpasses state-of-the-art techniques by a large margin, attaining 97% accuracy with only 20 parameters and keeping processing costs to a bare minimum, and Fig. 16 compares three ML methods. The comparative analysis of the proposed methodology with other ML and DL methods shows that the proposed methodology performed comparatively better with all factors (Table 7).

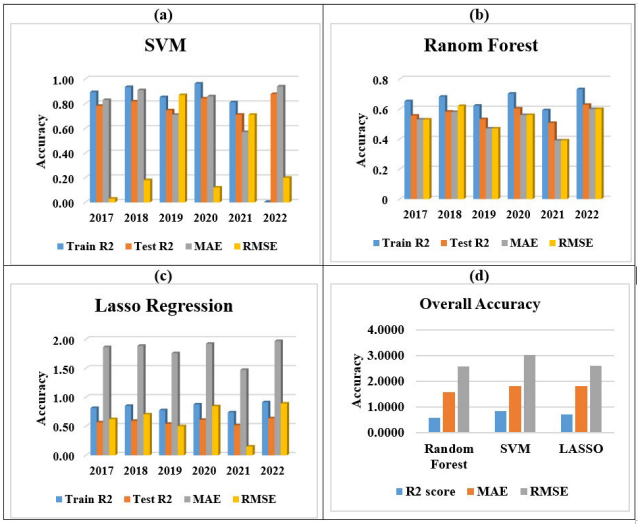

FIGURE 17. The modeling performance of the three methods (predicted R^2, RMSE, and MAE) is based on individual climate and remote sensing predictors during the growing season. A shows SVM accuracy, b shows RF accuracy, c shows LASSO accuracy, and d shows the three overall methods of predictor accuracy from 207 to 2022.

G. PERFORMANCE CAPABILITY OF ML ALGORITHMS TOWARD BETTER YIELD MODELING

Despite the variety of models used for agricultural output forecasting and prediction, their practicality is often compromised by including numerous factors and assumptions, making them less feasible compared to statistical or ML methods [49]. Throughout different phenological stages of crop growth, crop canopy variability is manifested distinctively in spectral reflectance bands. For instance, during the early to middle stages of wheat growth, biomass is reflected in the near-infrared band, while during the late growth stages, it increases in the red and blue bands. With the investigation of this study, the ML yield models, specifically Random Forest (RF), Support Vector Machine (SVM), and LASSO, demonstrated effective performance, as depicted in Table 6 and illustrated in Fig. 17 [54].

In assessing model performance using R^2, RMSE, MAE, and Random Forest (RF), the RF emerged as the top-performing model for yield prediction in this study, where both remote sensing and meteorological data were used. This superior performance aligns with previous studies on multi-grain yield estimates that leverage remote sensing and climatic data products. Moreover, applying ML techniques facilitated the practical exploration of non-linear interactions among various meteorological variables, such as temperature and rainfall. Despite outperforming non-linear ML algorithms in prediction accuracy, the linear Least Absolute Shrinkage and Selection Operator (LASSO) surpassed RF. Consequently, The findings highlight LASSO as the second most impactful model, following RF, regarding the coefficient of determination. Finally, upon comparing non-linear and linear yield prediction models, it was observed that both RF and LASSO emerged as significant techniques for yield prediction [55]. Linear models, such as Lasso and RF, performed well because their linear modeling structure cannot capture the nonlinear (Deep Learning) effects of weather and soil factors [56].

This study revealed that LASSO exhibited less robustness and computing time complexity than RF while still delivering high prediction accuracies when identifying linear correlations between explanatory and response variables. On the other hand, RF proved advantageous in detecting concealed non-linear interactions among variables during the peak growing season of wheat in the region [57] and demonstrated the highest accuracy in forecasting wheat production using NDVI and meteorological indicators [58]. Based on the research and findings, the suggested stack-ensemble-based wheat yield model is a useful and novel way to predict yields. The positive outcomes compared to the benchmark and other ML models and with other crops like rice confirm this conclusion. It is important to note that implementing these models at a larger regional or local field level may yield different outcomes, presenting an avenue for further exploration in future research.

V. CONCLUSION

In conclusion, this study aimed to identify the optimal model, data sources, and spectral band combination for winter wheat yield estimation using Landsat 8 remote sensing imagery, employing the RF, LASSO, and SVM ML approaches. Analyzing the temporal relationship between satellite data and winter wheat production revealed that, among the three models, the RF model outperformed SVM and LASSO, achieving an impressive accuracy of 97%. The integration of three ML techniques with climate and remote sensing data demonstrated the potential for more accurate yield predictions for the entire Multan district of Pakistan ($R^2 = 0.78 - 0.88$) compared to using either climate data alone ($R^2 = 0.65 - 0.732$) or remote sensing data alone ($R^2 = 0.49$-0.70). The study found that two non-linear techniques, notably those containing water-related predictors ($R^2 = 0.82 - 0.92$), outperformed temperature-related predictors ($R^2 = 0.34 - 0.63$) in terms of yield prediction within the winter wheat planting zone in Multan District. Furthermore, the individual predictive performance of the three models outperformed the traditional yield prediction approach in the winter wheat planting area of Multan, Punjab, Pakistan.

Furthermore, temperature data collected throughout the growth phase supplemented remote sensing data for yield prediction. However, including remote sensing data lowered the impact of meteorological data across time, from the beginning to the end of the growing season, on wheat production forecast. More trustworthy models and publically available data would be required to improve the precision of large-scale yield projections in the spring wheat planting zone. The scope of the study could be publicly expanded to include other provinces of tan and regions heavily dependent on agriculture, such as central Punjab and the Pothohar region. Future expansions involve exploring the impacts of floods, waterlogging, salinity, and other factors on the wheat crop yield.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] M. Marszalek, M. Körner, and U. Schmidhalter, "Prediction of multi-year winter wheat yields at the field level with satellite and climatological data," *Comput. Electron. Agricult.*, vol. 194, Mar. 2022, Art. no. 106777.

[2] 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.

[3] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop yield forecasting on the Canadian prairies by remotely sensed vegetation indices and machine learning methods," *Agricult. Forest Meteorol.*, vol. 218, pp. 74–84, Mar. 2016.

[4] R. Tanabe, T. Matsui, and T. S. T. Tanaka, "Winter wheat yield prediction using convolutional neural networks and UAV-based multispectral imagery," *Field Crops Res.*, vol. 291, Feb. 2023, Art. no. 108786.

[5] J. Dempewolf, B. Adusei, I. Becker-Reshef, B. Barker, P. Potapov, M. Hansen, and C. Justice, "Wheat production forecasting for Pakistan from satellite data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2013, pp. 3239–3242.

[6] T. Sadad, S. A. C. Bukhari, A. Munir, A. Ghani, A. M. El-Sherbeeny, and H. T. Rauf, "Detection of cardiovascular disease based on PPG signals using machine learning with cloud computing," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Aug. 2022.

[7] A. Badshah, A. Ghani, A. Daud, A. Jalal, M. Bilal, and J. Crowcroft, "Towards smart education through Internet of Things: A survey," *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–33, Feb. 2024.

[8] Y. Di, M. Gao, F. Feng, Q. Li, and H. Zhang, "A new framework for winter wheat yield prediction integrating deep learning and Bayesian optimization," *Agronomy*, vol. 12, no. 12, p. 3194, Dec. 2022.

[9] M. U. Ahmed and I. Hussain, "Prediction of wheat production using machine learning algorithms in northern areas of Pakistan," *Telecommun. Policy*, vol. 46, no. 6, Jul. 2022, Art. no. 102370.

[10] K. Lashari. (1974). *Land Use Atlas of Pakistan*. [Online]. Available: https://wedocs.unep.org/bitstream/handle/20.500

[11] D. Beillouin, B. Schauberger, A. Bastos, P. Ciais, and D. Makowski, "Impact of extreme weather conditions on European crop production in 2018," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 375, Oct. 2020, Art. no. 20190510.

[12] T. Ben-Ari, J. Boé, P. Ciais, R. Lecerf, M. Van Der Velde, and D. Makowski, "Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France," *Nature Commun.*, vol. 9, no. 1, p. 1627, Apr. 2018.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[14] L. Rattis, P. M. Brando, M. N. Macedo, S. A. Spera, A. D. A. Castanho, E. Q. Marques, N. Q. Costa, D. V. Silverio, and M. T. Coe, "Climatic limit for agriculture in Brazil," *Nature Climate Change*, vol. 11, no. 12, pp. 1098–1104, Dec. 2021.

[15] R. Buitenwerf, L. Rose, and S. I. Higgins, "Three decades of multi-dimensional change in global leaf phenology," *Nature Climate Change*, vol. 5, no. 4, pp. 364–368, Apr. 2015.

[16] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, K.-M. Shim, J. S. Gerber, V. R. Reddy, and S.-H. Kim, "Random forests for global and regional crop yield predictions," *PLoS ONE*, vol. 11, no. 6, Jun. 2016, Art. no. e0156571.

[17] G. Azzari, M. Jain, and D. B. Lobell, "Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries," *Remote Sens. Environ.*, vol. 202, pp. 129–141, Dec. 2017.

[18] 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.

[19] A. Kern, Z. Barcza, H. Marjanovic, T. Árendás, N. Fodor, P. Bónis, P. Bognár, and J. Lichtenberger, "Statistical modelling of crop yield in central Europe using climate data and remote sensing vegetation indices," *Agricult. Forest Meteorol.*, vol. 260, pp. 300–320, Oct. 2018.

[20] S. A. Shammi and Q. Meng, "Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling," *Ecological Indicators*, vol. 121, Feb. 2021, Art. no. 107124.

[21] C. Wu, J. Wang, P. Ciais, J. Peñuelas, X. Zhang, O. Sonnentag, F. Tian, X. Wang, H. Wang, R. Liu, Y. H. Fu, and Q. Ge, "Widespread decline in winds delayed autumn foliar senescence over high latitudes," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 16, Apr. 2021, Art. no. e2015821118.

[22] 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.

[23] V. S. Manivasagam and O. Rozenstein, "Practices for upscaling crop simulation models from field scale to large regions," *Comput. Electron. Agricult.*, vol. 175, Aug. 2020, Art. no. 105554.

[24] J. Cao, Z. Zhang, F. Tao, L. Zhang, Y. Luo, J. Han, and Z. Li, "Identifying the contributions of multi-source data for winter wheat yield prediction in China," *Remote Sens.*, vol. 12, no. 5, p. 750, Feb. 2020.

[25] B. Barnabás, K. Jäger, and A. Fehér, "The effect of drought and heat stress on reproductive processes in cereals," *Plant, Cell Environ.*, vol. 31, no. 1, pp. 11–38, Jan. 2008.

[26] N. T. Son, C. F. Chen, C. R. Chen, V. Q. Minh, and N. H. Trung, "A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation," *Agricult. Forest Meteorol.*, vol. 197, pp. 52–64, Oct. 2014.

[27] S. Data, "Crop yield estimation using decision trees and random forest machine learning algorithms on data from terra," *Mach. Learn. Data Mining Aerosp. Technol.*, vol. 836, p. 107, Jul. 2019.

[28] N.-T. Son, C.-F. Chen, Y.-S. Cheng, P. Toscano, C.-R. Chen, S.-L. Chen, K.-H. Tseng, C.-H. Syu, H.-Y. Guo, and Y.-T. Zhang, "Field-scale rice yield prediction from Sentinel-2 monthly image composites using machine learning algorithms," *Ecological Informat.*, vol. 69, Jul. 2022, Art. no. 101618.

[29] S. T. Arab, R. Noguchi, S. Matsushita, and T. Ahamed, "Prediction of grape yields from time-series vegetation indices using satellite remote sensing and a machine-learning approach," *Remote Sens. Appl., Soc. Environ.*, vol. 22, Apr. 2021, Art. no. 100485.

[30] N.-T. Son, C.-F. Chen, C.-R. Chen, H.-Y. Guo, Y.-S. Cheng, S.-L. Chen, H.-S. Lin, and S.-H. Chen, "Machine learning approaches for rice crop yield predictions using time-series satellite data in Taiwan," *Int. J. Remote Sens.*, vol. 41, no. 20, pp. 7868–7888, Oct. 2020.

[31] P. S. M. Gopal, "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms," *Appl. Artif. Intell.*, vol. 33, no. 7, pp. 621–642, Jun. 2019.

[32] Y. Ma, Z. Zhang, Y. Kang, and M. Özdogan, "Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach," *Remote Sens. Environ.*, vol. 259, Jun. 2021, Art. no. 112408.

[33] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Comput. Electron. Agricult.*, vol. 151, pp. 61–69, Aug. 2018.

[34] L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis, "Machine learning in agriculture: A comprehensive updated review," *Sensors*, vol. 21, no. 11, p. 3758, May 2021.

[35] P. Q. Khang, K. Kaczmarczyk, P. Tutak, P. Golec, K. Kuziak, R. Depczynski, M. Hernes, and A. Rot, "Machine learning for liquidity prediction on Vietnamese stock market," *Proc. Comput. Sci.*, vol. 192, pp. 3590–3597, Jan. 2021.
[36] 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.
[37] D. Gouache, A.-S. Bouchon, E. Jouanneau, and X. Le Bris, "Agrometeorological analysis and prediction of wheat yield at the departmental level in France," *Agricult. Forest Meteorol.*, vols. 209–210, pp. 1–10, Sep. 2015.
[38] N. Hanasaki, S. Fujimori, T. Yamamoto, S. Yoshikawa, Y. Masaki, Y. Hijioka, M. Kainuma, Y. Kanamori, T. Masui, K. Takahashi, and S. Kanae, "A global water scarcity assessment under shared socio-economic pathways—Part 2: Water availability and scarcity," *Hydrol. Earth Syst. Sci.*, vol. 17, no. 7, pp. 2393–2413, Jul. 2013.
[39] J. F. Progga, M. N. H. Khan, and M. G. M. Amin, "Meteorological parameters-soil temperature relations in a sub-tropical summer grassland: Physically-based and data-driven modeling," *Ataturk Univ. J. Agricult. Fac.*, vol. 54, no. 2, pp. 48–56, May 2023.
[40] S. Hussain, M. Mubeen, A. Ahmad, N. Masood, H. M. Hammad, M. Amjad, M. Imran, M. Usman, H. U. Farid, S. Fahad, W. Nasim, H. M. R. Javeed, M. Ali, S. A. Qaisrani, A. Farooq, M. S. Khalid, and M. Waleed, "Satellite-based evaluation of temporal change in cultivated land in Southern Punjab (Multan region) through dynamics of vegetation and Land Surface Temperature," *Open Geosci.*, vol. 13, no. 1, pp. 1561–1577, Dec. 2021.
[41] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[42] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep CNN-LSTM model," *Sensors*, vol. 19, no. 20, p. 4363, Oct. 2019.
[43] P. Lemenkova and O. Debeir, "Multispectral satellite image analysis for computing vegetation indices by R in the Khartoum region of Sudan, northeast Africa," *J. Imag.*, vol. 9, no. 5, p. 98, May 2023.
[44] 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.
[45] U. Hayat, S. Ali, A. Mateen, and H. Bilal, "The role of agriculture in poverty alleviation: Empirical evidence from Pakistan," *Sarhad J. Agricult.*, vol. 35, no. 4, pp. 1309–1315, 2019.
[46] M. J. Roberts, N. O. Braun, T. R. Sinclair, D. B. Lobell, and W. Schlenker, "Comparing and combining process-based crop models and statistical models with some implications for climate change," *Environ. Res. Lett.*, vol. 12, no. 9, Sep. 2017, Art. no. 095010.
[47] E. Kamir, F. Waldner, and Z. Hochman, "Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods," *ISPRS J. Photogramm. Remote Sens.*, vol. 160, pp. 124–135, Feb. 2020.
[48] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[49] H. Khan and S. Ghosh, "Machine learning approach for crop yield prediction emphasis on K-medoid clustering and preprocessing," in *Proc. Int. Conf. Intell. Comput. Smart Commun.* Springer, 2020, pp. 287–299.
[50] K. Murakami, S. Shimoda, Y. Kominami, M. Nemoto, and S. Inoue, "Prediction of municipality-level winter wheat yield based on meteorological data using machine learning in Hokkaido, Japan," *PLoS ONE*, vol. 16, no. 10, Oct. 2021, Art. no. e0258677.
[51] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: A big comparison for NAS," 2019, *arXiv:1912.06059*.
[52] S. K. Sahu, A. Mokhade, and N. D. Bokde, "An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges," *Appl. Sci.*, vol. 13, no. 3, p. 1956, Feb. 2023.
[53] F. Rundo, F. Trenta, A. L. Di Stallo, and S. Battiato, "Machine learning for quantitative finance applications: A survey," *Appl. Sci.*, vol. 9, no. 24, p. 5574, Dec. 2019.
[54] S. Arshad, J. H. Kazmi, M. G. Javed, and S. Mohammed, "Applicability of machine learning techniques in predicting wheat yield based on remote sensing and climate data in Pakistan, South Asia," *Eur. J. Agronomy*, vol. 147, Jul. 2023, Art. no. 126837.
[55] L. D. Estes, B. A. Bradley, H. Beukes, D. G. Hole, M. Lau, M. G. Oppenheimer, R. Schulze, M. A. Tadross, and W. R. Turner, "Comparing mechanistic and empirical model projections of crop suitability and productivity: Implications for ecological forecasting," *Global Ecol. Biogeography*, vol. 22, no. 8, pp. 1007–1018, Aug. 2013.
[56] A. K. Srivastava, N. Safaei, S. Khaki, G. Lopez, W. Zeng, F. Ewert, T. Gaiser, and J. Rahimi, "Winter wheat yield prediction using convolutional neural networks from environmental and phenological data," *Sci. Rep.*, vol. 12, no. 1, p. 3215, Feb. 2022.
[57] Y. Cai, K. Guan, D. Lobell, A. B. Potgieter, S. Wang, J. Peng, T. Xu, S. Asseng, Y. Zhang, L. You, and B. Peng, "Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches," *Agricult. Forest Meteorol.*, vol. 274, pp. 144–159, Aug. 2019.
[58] L. Li, B. Wang, P. Feng, D. Li Liu, Q. He, Y. Zhang, Y. Wang, S. Li, X. Lu, C. Yue, Y. Li, J. He, H. Feng, G. Yang, and Q. Yu, "Developing machine learning models with multi-source environmental data to predict wheat yield in China," *Comput. Electron. Agricult.*, vol. 194, Mar. 2022, Art. no. 106790.

MUHAMMAD ASHFAQ received the M.S. degree in software engineering from International Islamic University Islamabad, Pakistan, in 2012. He is currently a Ph.D. Scholar in software engineering with International Islamic University Islamabad. His research interests include software development, machine learning, deep learning, and artificial intelligence.

IMRAN KHAN received the M.C.S. and M.S. degrees in computer science from International Islamic University Islamabad, Pakistan, in 2002 and 2005, respectively, and the Ph.D. degree from the Department of Computer Science, International Islamic University Islamabad, in 2017. His academic journey commenced with International Islamic University Islamabad. These formative years laid the groundwork for his expertise and commitment to the field. He is currently an Assistant Professor with the Department of Computer Science, International Islamic University Islamabad. His active role in academia is complemented by his contributions to various ICT research and development-funded projects at the same institution, showcasing his dedication to advancing knowledge and technology. His research interests include reflecting his profound curiosity and passion for the field, information security and privacy, exchange and medical data management, natural language processing, next-generation networks, and cloud computing. His contributions in these domains exemplify his commitment to shaping the future of computer science and technology.

ABDULRAHMAN ALZAHRANI is currently an Assistant Professor with the Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia. He is also the Head of the Computer Network Department. His research interests include information technology and innovations in data science, machine learning, and health informatics. He teaches several courses to bachelor's and master's students. He teaches e-commerce and data visualization classes.

MUHAMMAD USMAN TARIQ (Member, IEEE) is currently an Associate Professor with Abu Dhabi University, Abu Dhabi Campus. Previously, he has held various academic and leadership appointments over the past 15 years in industry and academia. He has been a consultant, a master trainer, and an assessor for local and international organizations. He has diverse and significant experience working with ABET, ACBSP, AACSB, WASC, CAA, EFQM, AMBA, BGA, and NCEAC accreditation agencies. Additionally, he has operational expertise in incubators, research laboratories, government research projects, case centers, private sector startups, program creation, and management at various industrial and academic levels. He is a Certified Higher Education Teacher with Harvard University, USA, a Certified Change Leader, an Assessor with EFQM Brussels, CIPD Level 7 SRLD Certified, a Certified Online Educator with HMBSU, a Certified Six Sigma Master Black Belt, and a Lead Auditor of ISO 9001 Certified, ISO 14001, IOSH MS, OSHA 30, and OSHA 48. He has been awarded a Certified Chartered Fellowship from CIPD U.K. He holds professional memberships with IEEE, APPAM, and ACM.

HUMERA KHAN received the M.S. and M.B.A. degrees from International Islamic University Islamabad, in 2006 and 2009, respectively, and the Ph.D. degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2012. She is currently an Assistant Professor with the Department of Information Systems, College of Computing and Information Technology, Rafha. She joined Northern Border University, Rafha, in 2016. She has 15 years of extensive teaching and industry experience in marketing and management. She worked in United Arab Emirates, Malaysia, and Pakistan and taught various business and IT subjects in marketing, management, and information systems. Her research interests include digital marketing, technology acceptance, e-commerce, e-business, and online consumer behavior. She is also working on the quality committee of ABET and NCA accreditation for the College of Computing and Information Technology, Northern Border University.

ANWAR GHANI received the B.S. degree in computer science from the University of Malakand, Khyber Pakhtunkhwa, Pakistan, in 2007, and the M.S. and Ph.D. degrees in computer science from the Department of Computer Science and Software Engineering, International Islamic University Islamabad, in 2011 and 2016, respectively. He is a Post-Doctorate Fellow at the Big Data Research Center, Jeju National University, South Korea. He also holds a faculty position within the Department of Computer Science, International Islamic University, Islamabad. Before starting his academic career, he gained practical experience as a Software Engineer at Bioman Technologies from 2007 to 2011. His dedication to academic excellence led to his selection as an exchange student under the EURECA program in 2009, allowing him to study at VU University, Amsterdam, The Netherlands. He continued to expand his academic horizons through the EXPERT program in 2011, allowing him to study at Masaryk University in the Czech Republic, with funding provided by the European Commission. His research interests encompass several domains, including wireless sensor networks, Information Security, Internet of Things, and Edge Computing.